COMPREHENSIVE IN SILICO ASSESSMENT OF CANCER CELL LINE FIDELITY

by

Pavithra Kumar

A thesis submitted to The Johns Hopkins University in conformity with the requirements for the degree of Master of Science and Engineering in Biomedical Engineering.

Baltimore, Maryland

May 2017

Abstract

Since 1971, when US president Richard Nixon declared a "war against cancer", an estimated $200 billion has been spend on research and drug development for cancer. Globally, the cancer incidence and the mortality rates have declined over the past decade, owing to progress made in healthcare and biological research. In order to study the biological properties of cancers and to enable novel drug discovery, in-vitro model systems are used and the most common amongst these are cancer cell lines. These are primary-tumor derived cells that are propagated in the lab, and are used to study both the biological and pharmacological aspects of cancer. The use of cell lines in research, in the field of cancer diagnostics, has resulted in a number of drugs that are now available in the market to combat cancers.

Since the isolation of the first cancer cell line, HeLa, in 1951, a number of cell lines are available for various types of cancers and these have been around for decades now. Cell lines in culture not only lose the properties of their cells of origin gradually, they also acquire mutations over each passage that causes the lines to behave very differently as compared to primary tumors. Additionally, there is competition among populations of cells in culture and it is widely believed that cell lines that are presently available are largely clonal populations of those cells that grow faster and survive better as compared to others. This results in loss of heterogeneity that is characteristic of in-vivo cancers.

In order to quantitatively assess the fidelity of these cell lines, we use a modified version of the network biology platform CellNet to construct tumor type specific gene regulatory networks for twelve common primary cancers and train random forest classifiers. Using expression data from 917 cell lines as the query and comparing their GRNs (Gene Regulatory Networks) against the tumor type specific classifiers, the probability scores for each

of these cell lines to resemble primary tumors are calculated. Those cell lines that show a high classification score to their tumor type of origin are considered ideal in vitro models for that cancer. The GRN status and network influence scores were also calculated for cell lines and using these statistics and CellNet predictions, in-vitro models for cancer research are suggested.

Additionally, we propose the use of Cancer CellNet as a clinical diagnostic platform to resolve the tumor identity in unknown cases and demonstrate its ability to obtain this information using tumor samples where the cells of origin of cancer remain unknown.

Advisors: Patrick Cahan[1], Aleksander Popel[2], Winston Timp[3]

[1]Primary advisor: Assistant Professor, Department of Biomedical Engineering, Institute of Cell Engineering, Johns Hopkins University.

[2]Professor, Department of Biomedical Engineering, School of Medicine, Johns Hopkins University.

[3]Assistant Professor, Department of Biomedical Engineering, John Hopkins University.

Acknowledgements

I would like to take this opportunity to thank Dr. Patrick Cahan for being an amazing mentor, and for making this Hopkins journey, a dream come true. His high expectations have pushed me to become a better scientist and his enormous patience, impeccable work ethic and wonderful personality have set very high standards that I hope to reach one day. Working with him has truly been a pleasure. I would also like to thank Dr. Aleksander Popel and Dr. Winston Timp for reading through my work and providing useful and timely inputs.

I want to thank Kathleen and Edroaldo for doing the initial analysis for this project and my lab mates, Remy, Emily, Abby, Yuqi, Qin and Emily Su for supporting me through everything, teaching me to code and for all the memories that include shopping, deep emotional conversations and watching movies in the lab. I also want to thank Segun, Debangshu and all members of the Semenza lab for their company, conversations and all their help while we were setting up the lab.

My Hopkins journey would have been incomplete without Brant, Alexis and Vijay; thank you for all the trips, parties, board game nights and adventures. I also want to thank my Shakti family for sharing my passion for dance and pizza; I truly love each and every one of you girls and the absolute joy that I get every time I dance with you, both at practice and during competitions, is unparalleled.

Lastly, I wish to acknowledge my family, without whose support, none of this would have been possible. Thank you amma, appa, Raghav, athai, athimber and Dhaaru for supporting me through my decision of coming to Hopkins and every single step of the way after that. I am who I am because of you and I am forever grateful; Sridhar Chitapa and Chithi, thank you so much for showing me unconditional love, encouragement and support; Apoorva, thanks for being the sister that I never had and finally, thank you Kaushik, for being an amazing companion, and for your blind faith and confidence in me. From editing my Hopkins application to editing my thesis, you have always been there for me, and I am truly grateful.

Thank you all so much!!!!

Dedication

To all those fighters and survivors out there.....

.....who believed that there is a "Can" in Cancer!

Table of Contents

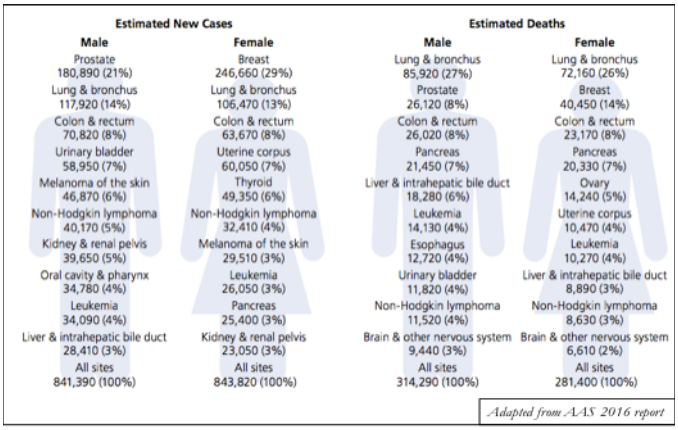
List of Figures

List of Tables

Chapter 1 – Introduction

1.1 Motivation for the project

The American Cancer Society estimated a total of 14.5 million people living with a history of cancer in the United States of America on January 1 2014 and by 2017, a cumulative total of 1.7 billion new cases (1930-2017) are expected to be diagnosed globally. Around 35% of this number are expected to succumb to this deadly disease by 2016, which translates to approximately 1650 deaths a day[1,2]. This makes cancer the second deadliest disease, next only to cardiovascular disease. The sites where cancer is predicted to occur in these cases, and the projected percentage of deaths is depicted in Figure 1.1.



Estimated New Cases		Estimated Deaths	
Male	**Female**	**Male**	**Female**
Prostate 180,890 (21%)	Breast 246,660 (29%)	Lung & bronchus 85,920 (27%)	Lung & bronchus 72,160 (26%)
Lung & bronchus 117,920 (14%)	Lung & bronchus 106,470 (13%)	Prostate 26,120 (8%)	Breast 40,450 (14%)
Colon & rectum 70,820 (8%)	Colon & rectum 63,670 (8%)	Colon & rectum 26,020 (8%)	Colon & rectum 23,170 (8%)
Urinary bladder 58,950 (7%)	Uterine corpus 60,050 (7%)	Pancreas 21,450 (7%)	Pancreas 20,330 (7%)
Melanoma of the skin 46,870 (6%)	Thyroid 49,350 (6%)	Liver & intrahepatic bile duct 18,280 (6%)	Ovary 14,240 (5%)
Non-Hodgkin lymphoma 40,170 (5%)	Non-Hodgkin lymphoma 32,410 (4%)	Leukemia 14,130 (4%)	Uterine corpus 10,470 (4%)
Kidney & renal pelvis 39,650 (5%)	Melanoma of the skin 29,510 (3%)	Esophagus 12,720 (4%)	Leukemia 10,270 (4%)
Oral cavity & pharynx 34,780 (4%)	Leukemia 26,050 (3%)	Urinary bladder 11,820 (4%)	Liver & intrahepatic bile duct 8,890 (3%)
Leukemia 34,090 (4%)	Pancreas 25,400 (3%)	Non-Hodgkin lymphoma 11,520 (4%)	Non-Hodgkin lymphoma 8,630 (3%)
Liver & intrahepatic bile duct 28,410 (3%)	Kidney & renal pelvis 23,050 (3%)	Brain & other nervous system 9,440 (3%)	Brain & other nervous system 6,610 (2%)
All sites 841,390 (100%)	All sites 843,820 (100%)	All sites 314,290 (100%)	All sites 281,400 (100%)

Adapted from AAS 2016 report

Figure 1.1 – Gender-wise estimates for leading sites of new cancer cases (2016)[1]

Even though the number of people suffering from cancer seems to be an appalling number, the number of cancer deaths has reduced by 23% from the projected numbers, due to advances in diagnosis and treatment, as well as significant improvements in lifestyle, thus avoiding almost 1.7 million cancer deaths. A trend line showing the year-wise mortality rates per cancer-type for both males and females, from 1930 – 2012 is shown in Figure 1.2, and it shows the decline in the number of deaths since 1990[1].

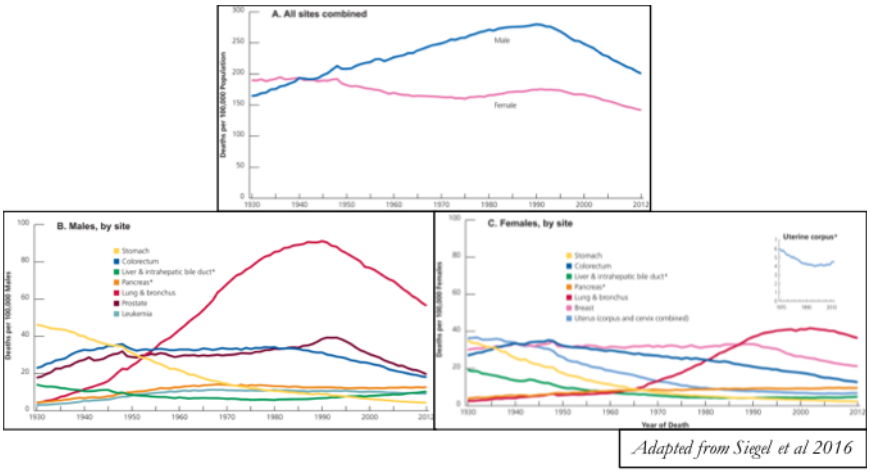


Figure 1.2 – Gender-wise cancer mortality rates by site from 1930-2012 in the USA[3]

Advances made in biomedical research has contributed directly in this decline in mortality rates due to cancer, not only by providing therapeutic drugs that help to cure or contain cancer but also by improving diagnostic tools that aid in early identification of the disease, thus improving the disease prognosis. The National Cancer Institute is one of the biggest funding bodies in America supporting investigator-driven cancer research and in 2017, an estimated $5.21billion in total has been allocated for this purpose.

Human-cancer derived cell lines are the most popular cancer models used in research and they are used not only for advancing our understanding of the disease but also for developing novel therapeutics and improving efficacy of treatment[4]. For decades, cell lines lay the foundation for advancement in cancer biology and many models were established for studying various cancers. However, there was always criticism in using cell lines as cancer models as there are modifications that are introduced in cells isolated from tumors to enable their propagation indefinitely in culture. Thus, a lot of the cell lines do not actually recapitulate the features of tumors in vivo. For example, glioblastoma-derived cell lines form a ball of cells as opposed to a spider-web like infiltration as observed in human brains and thus, do not make accurate models for the primary tumor[5]. However, there are other cell lines models that have

proven instrumental in the development of drugs. An example of this would be the MCF-7 breast cancer cell line that has been used in clinical trials for many drugs like Faslodex, which have been established as successful drugs in the market[6]. Cell lines are also useful to study rare cancers, where primary samples cannot be obtained readily due to the complex location of the tumor or its rarity among the population[7].

There are other concerns regarding the use of cell lines as tumor analogs, like the contamination of cell lines with cells from other cell lines (cross-contamination). Back in the 1970s, researchers discovered through cytogenetic studies that cell lines used as models for various cancers were all derived from the HeLa cell line, which was a cervical cancer cell line[8] and not the tumors they intended them to represent. Immortalizing primary cells obtained from humans in order to convert them into a cell line has also been shown to modify the characteristics of the cells, deviating from the exact tumor profile.

With advances in genomic technologies, many primary tumor samples have been sequenced and genomic information is now available, the largest of which is the Cancer Genome Atlas (TCGA)[9]. Back to back publications in 2010 provide a comprehensive analysis of mutations and copy number variations (CNVs) in human cancers[10,11]. There was a debate for many years concerning the merit of sequencing cancer cell lines, but a stark reduction in the cost of genomics resulted in the publication of the Cancer Cell Line Encyclopedia (CCLE), a comprehensive study encompassing expression profiles of ~1000 cancer-derived cell lines[12].

A pilot study in 2013 by Domcke et al, compared genetic and expression profiles of the CCLE ovarian cancer cell lines, with profiles of primary tumors. The authors found that cell line models for ovarian cancer were those that showed least similarity to the in vivo tumor profile[13]. As part of their study, they also ranked ovarian cancer cell lines by their resemblance

to the ovarian tumors, so that ideal in vitro ovarian cancer models can be used in research.

In this thesis, I aim to assess the fidelity of vitro models for some of the most widely occurring cancers like breast cancer, prostate cancer and leukemia, among others. However, as opposed to comparing lines and tumors by global expression data, I use Gene Regulatory Network (GRNs)[14] as a basis for comparison. These GRNs are compared with those obtained from cell lines to classify them on the basis of their resemblance to those primary tumors whose GRN profile they most resemble. In addition to assessing cell line models for multiple cancers, this study will also explore possible molecular contributors differentiate between tumors and cell lines. The information obtained from analyzing thousands of primary tumor samples also paves way to explore the cellular origins of CUP (Cancers of Unknown Primary)[15] and Merkel Cell Carcinoma[16], which remains unknown.

1.2 Organization of the thesis

Chapter 1 highlights the motivation for this study. Chapter 2 provides the relevant literature review of the field and defines important terms that are used throughout this thesis. Chapter 3 contains the detailed workings of the computational protocol with the syntax expanded and the tools that were employed to run the code effectively. Chapter 4 describes the results obtained at the end of our computational analysis of various cell lines and Chapter 5 contains a comprehensive list of all the references used in this thesis. The entire Cancer CellNet code is provided as part of the Appendix.

Chapter 2 – Background

2.1 In vitro cancer models

Cancer is a collection of many diseases into one due to diverse cells of origin and plethora of molecular mechanisms underlying disease, not to mention the very strong environmental contribution to the triggering of certain types of cancers[17]. The genetic and environmental influence in determining the susceptibility of a person to get a positive cancer diagnosis makes cancer a complex disease, as opposed to being a metabolic or a genetic disorder. This molecular heterogeneity in the origin and propagation of cancer is what makes cancer one of the deadliest diseases in history and one of the hardest to cure.

In order to design molecules that can treat cancerous cells specifically, a thorough understanding of the cellular and molecular aspects of the disease needs to be achieved. The advances made in the field of cancer pathophysiology rely heavily on the availability of a variety of in-vitro model systems, each of which models a particular aspect of this diverse disease. A few such models are cancer cell lines[18], tumor primary cultures, primary tumors[19], xenografts and genetically modified mice[20]. The choice of cancer model depends on the aim of each particular study as each model system mentioned above is accompanied by its own set of advantages and limitations. Cancer models also provide a low cost screening platform for novel cancer therapeutic drugs.

The need for improved accuracy and relevance of tumor models has resulted in an increase in complexity of in vitro tumor models, where they progressed from just simple cell proliferation and cytotoxicity screens to those that incorporate metastasis, matrix remodeling and angiogenesis. This was made possible due to advances in the fields of tissue engineering, biomaterials, microfluidics and tumor biology[21]. There is also scope for the development of

patient-specific in vitro cancer models[22] in order to personalize the treatment regimen in certain rare or complex cases.

In vitro tumor models that currently exist in the field are designed to recapitulate not only the biological profile of cancers with respect to cell composition, but also the physical aspects like ECM structure, migration of cells, invasion, extravasation and angiogenesis. These are popularly referred to as 3-D tumor models and are used to study drug delivery mechanisms as well as tumor microenvironment complexity studies. The source of cancer cells for these models can be either cell lines or cells from the primary tumor itself. The latter source provides various advantages such as understanding the heterogeneity in tissue composition of the cancer, comparison of results across various biological samples and more importantly, recapitulating the in vivo tumor characteristics.

Advances made in diverse fields like microfluidics, fluid dynamics, biomaterials and polymers have led to the development of in vitro cancer models with increased complexity. However, currently available cancer models are classified broadly as the following: Transwell-based, Spheroid-based, Tumor-microvessel or hybrid model systems and these are explained in Katt et al 2016[21] in detail.

Despite these more advanced models, cell lines are by far the most widely used cancer models due to their low cost, ease of isolation and manipulation.

2.2 Cancer cell lines as tumor models

HeLa was the first cancer-derived cancer cell line to be established. It was derived from Henrietta Lacks, a cervical cancer patient at Johns Hopkins Hospital in 1951 by George Gay[23]. This cell line had infinite proliferative potential and was widely distributed across the world. Since the successful establishment of this cell line, similar culturing techniques were

used to establish cell lines from thousands of primary tumors over a period of two decades. These cell lines paved way to significantly advance the frontiers of cancer research, clinical diagnosis and development of therapeutic drugs. The major advantage cancer cell lines provide researchers is that they provide theoretically infinite supply of a relatively homogenous supply of cancer cells, to be experimented upon. A more detailed review of cancer cell lines and their uses in the field of drug development is provided by Ferriera et al[24]. The major disadvantages provided by cancer cell lines as an experimental model for tumors is listed below:

i. Genomic instability due to repeated culturing
ii. Dominance of specific clones over the other cell types that were originally a part of the tumor sample
iii. Cross contamination with other fast growing cell lines like HeLa
iv. Differences in environment of cells as compared to in vivo tumors
v. Loss of cellular heterogeneity in culture
vi. Bacterial and mycoplasma contamination
vii. Change in gene expression pathways due to absence of ECM signals present in tumor

There are over 3000 established cancer cell lines, isolated from multiple tissues and species. Characterization and re-authentication of these cell line models becomes crucial due to the following reasons: First is the high probability of cross contamination between cell lines due to their explosive proliferative potential[25,26] and second is the absence of advanced molecular techniques during the time of their initial isolation that could have resulted in an erroneous characterization of cell lines[4,13]. Another major factor is the passage number of these cell lines in culture as mutations tend to accumulate in cells that have been repeatedly cultured, thus deviating from the molecular profile of the parent tumor.

An example to illustrate the importance of accurate characterization of cell lines is the article published in 1999 that re-designated the cell line SW626 to be of colon origin as opposed to ovarian cancer which it was perceived to be since 1974[27]. Groups establishing cell lines are required to provide multiple lines of evidence to confirm the tumor type and also the cell-type of origin for the cell line[28], but such stringency was not possible during the early 1970s which was when a lot of the presently used cell lines were established. Hence, there is an active effort by the scientific community to redefine established cancer cell line models with modern molecular techniques[29], to re-instate the validity of these models as accurate representations of tumors for research. The various features that need to be characterized while establishing a new cell line are discussed in detail here[30].

As mentioned in section 1.1, Domcke et al initiated the efforts in sifting through all available cell line models available for ovarian cancers with an aim to identify the most accurate in vitro model for this cancer type. Although they did selective screening with parameters like CNVs and mutations in specific genes to compare tumors and cell lines, their results indicated the level of disparity in the popularity of cell line models and the extent of their resemblance to their parent tumor. Studies aiming to understand tumor characteristics or pre-clinical trial studies with cell lines need to ensure thorough characterization of the cell line prior to use as a tumor analog in the laboratory and this project provides one such metric to perform this comparison, in the form of GRN reconstruction and comparison between cells and tumors.

2.3 Gene Regulatory Networks (GRNs)

Cells are surrounded by extracellular matrix (ECM) that contains a myriad of signals which cells interpret and respond to accordingly. These signals vary depending on the environment and the physiological state of the organism and the responses are also

correspondingly different and they also vary between different cell types. This signal perception and response is achieved due to certain regulatory mechanisms that are programmed within the cell and these are executed by various proteins. One such regulatory mechanism is gene regulation and this is orchestrated by a number of proteins within the cell. Gene regulation can include both transcriptional and post-translational regulation, where the former is governed predominantly by transcription factors (TFs) and the latter by RNA binding proteins.

GRNs are mostly discussed in the context of transcriptional regulation. GRNs are essentially circuits that describe physical and/or regulatory interactions between transcription factors and their target genes[14]. GRNs include both physical interactions such as that of a TF with the DNA binding element or regulatory (logical) interactions which are more indirect in nature, and are inferred due to genetic experiments.

Figure 2.1 shows a schematic depicting the major elements that are a part of a transcriptional Gene Regulatory Network.

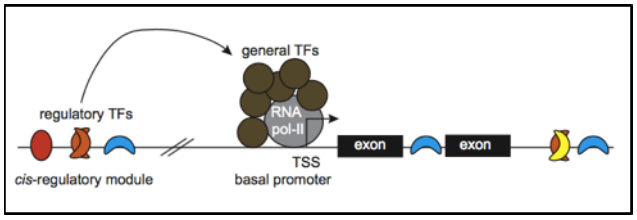


Figure 2.1 – Elements of a transcriptional GRN

Since transcriptional GRNs are essentially like circuits, they contain two elements: nodes and edges, where nodes are TFs and the target genes while edges are the regulatory relationship between them, which can be physical or logical. For a given cell in a given condition and a given environment, the GRN can vary anywhere from a few TFs and its corresponding targets to a complex mesh of multiple TFs, whose targets can be other TFs and thus, leading to a big network map, like the one shown in Figure 2.2[31].

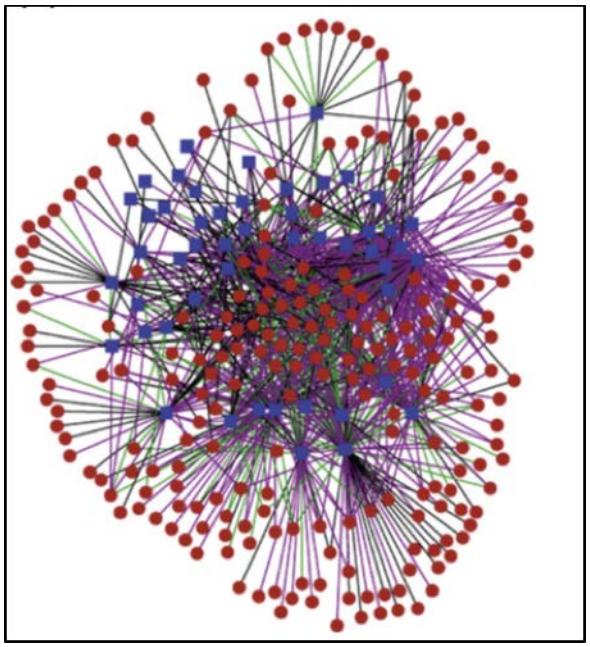


Figure 2.2 – GRN of interactions between *C.Elegans* TFs and metabolic target genes

Unlike gene expression profiles obtained from microarray platforms, GRNs provide a more comprehensive idea about the transcriptional state of a cell at a given time, as they not only provide information about TF expression, but also about their target gene regulation[32].

2.4 CellNet

CellNet is a network biology platform that was designed with the aim of assessing cell populations obtained through directed differentiation of induced pluripotent cells (iPSCs) or direct conversion of one cell type to another. It also suggests hypotheses aimed at improving these conversion protocols[33]. It uses a modified version of the CLR algorithm (refer section 2.5) to reconstruct cell-type specific GRNs. Once the classifiers are built, it quantitates the classification score of a given query sample into a particular tumor type using a Random Forest Classifier (refer section 2.6).

The pipeline takes as input, expression data obtained from public sources and using constructed cell/tissue type specific GRNs from hundreds of input sample of that particular

tissue as a basis for comparison, provides three outputs:

i. A classification probability score which indicates the extent to which the input has a profile indistinguishable in its expression profile as compared to a particular cell/tissue (classification score).
ii. The extent to which a cell/tissue specific GRN is established (GRN status)
iii. The list of TFs that are a part of the cell/tissue GRN, scored according to how dysregulated they are and by the dysregulation of their targets (Network Influence Score or NIS).

CellNet was initially designed to take as input, microarray expression data from mouse and human samples, and assess the extent of cell fate engineering from iPSCs. However, recently this pipeline has been modified to take in RNA-sequencing data as input, as RNA-sequencing is replacing microarrays as one of the cheapest genomic tools for whole transcriptome assessment and provides a more comprehensive view of the genetic state of a cell[34]. Until now, cell/tissue GRNs were constructed by CellNet using normal cells obtained from various tissue types as input and in this project, we modify the pipeline to construct tumor-type specific GRNs (ttGRNs) by providing primary tumor data as input (Cancer CellNet).

2.5 CLR algorithm

The Context Likelihood of Relatedness (CLR) algorithm is an unsupervised network inference method [35] and is an extension of the relevance inferences algorithm for the identification of transcriptional regulatory interactions[36,37]. The relevance method scored interactions between two genes on the basis of similarity in their expression levels and the threshold they set for extent of similarity determines the size of the network, as a lower

threshold will capture more interactions, some of which may be false positive and others which may be novel. CLR is similar to the relevance network but it adds another layer of complexity, by comparing the expression similarity of a particular interacting pair with the extent of similarity of other background interactor pair. Thus, it puts each pair in context with others, and eliminates all those interactions that are weak or non-specific, thus eliminating a huge percentage of the false positives that are captured by the relevance network. CLR can achieve a high-precision regulatory interaction network with as few as 60 expression datasets as input, and this becomes a lower threshold for training CellNet's pipeline.

2.6 Random Forest classifiers

Random Forests are an ensemble learning method that operates by constructing multiple decision trees and providing as output, either the mean prediction of regression of each individual tree or a classification score that is a mode of the individual classes[38]. In GRN reconstruction, the decision trees start with one gene that is either randomly selected or is selected based on the importance of a gene to that particular cell type.

CellNet employs random forest classifiers to construct the ttGRN based classifiers which will serve as the basis for comparison with the query datasets. For each platform, a single RF classifier was trained on a particular tumor type. Every gene that is specific to the ttGRN is used as a predictor variable in the construction of these trees, and the output of each tree is binary. GRN status of a sample can sometimes be weighed by the importance of a gene to the Random Forest classifier, as opposed to the mean expression profile in the training tumor type.

Chapter 3 – Materials & Protocol

In this chapter, the computational tools used for analyzing the 917 cancer cell lines from the Cancer Cell Line Encyclopedia is described in brief. The cancer CellNet pseudo code is presented, in detail,

The complete Cancer CellNet code is available in the Appendix.

3.1 Using Amazon Web Services (AWS)

Storage of the expression data files for all the samples to be queried and executing Cancer CellNet was achieved using Amazon Web Services (AWS), which is a comprehensive cloud computing platform. Among the growing selection of services made available for users by AWS, this project utilized EC2 and S3. EC2 or Elastic Compute Cloud allows customers to run applications on virtual interfaces called instances, which can be configured by the user and are theoretically unlimited in number. S3 or Simply Storage Service is a scalable, high speed, low cost storage service intended to backup or archive user data. The various steps involved in using this system is described below in detail.

3.1.1 Launching an EC2 instance

An instance is the virtual server in Amazon's Elastic Compute Cloud (EC2), which is used to run applications and computations on the AWS infrastructure. The configuration of the virtual interface used to perform the desired function can be assembled from a variety of instance types, each with different memory, storage capacity and computing power to suit the needs of the consumer.

Instances are created from Amazon Machine Images or AMI's, which are essentially templates that are configured with an Operating System and software in accordance with the needs of the user. A variety of in-build AMIs are made available for use but users can also

create their own AMIs, by customizing the virtual machine or VM to determine the working environment. For this project, the AMI called CellNet_RNA-seq_Pub (AMI ID - ami-1ad4430d) was launched with c3.8x large instance type and launch-wizard-1 security group.

Once an instance is launched and is running successfully, the terminal application is used to securely connect to the virtual machine by using the following command:

```
ssh –i aws_private_key_path ec2-user@instance_public_dns
```

The private key pair is locally downloaded and saved in your working directory and is detected when one attempts to connect to an instance. Once the connection is successful and you enter the virtual machine, all computations are performed in the cloud

3.1.2 Data handling in S3

Amazon S3 is an object storage device, which is different as compared to a block storage or a file storage device. The S3 cloud service provides the user with services similar to the one used by Amazon websites, which includes the facility to upload, download and store any object in its cloud database, up to a 5GB in size. The S3 cloud is divided into logical units of storage called buckets which are used for storing objects, consisting of data plus the metadata that is used to describe the data.

3.1.3 AWS Command Line Interface (CLI)

The CLI is a tool that helps manage AWS services by automating through scripts and help controlling the AWS through the command line. It includes a set of file commands that make transfers to and from S3 highly efficient and can be achieved in parallel. The commands that are used as a part of the Cancer CellNet code is described in detail below:

<u>To view the contents of a bucket/folder in S3:</u>

```
$ aws s3 ls s3://mybucket
$ aws s3 ls s3://mybucket/folder_path
```

To copy contents of a local folder to a folder in S3 or vice-versa:

```
    $ aws s3 cp myfolder s3://mybucket/myfolder --recursive

 upload: myfolder/file1.txt to s3://mybucket/myfolder/file1.txt
 upload: myfolder/file2.txt to s3://mybucket/myfolder/file2.txt

    $ aws s3 cp s3://mybucket/myfolder myfolder --recursive

download: s3://mybucket/myfolder/file1.txt to myfolder/file1.txt
download: s3://mybucket/myfolder/file2.txt to myfolder/file2.txt
```

The --recursive tag is used to ensure that all the contents within a folder are copied into the destination folder.

To sync the contents of a local folder to a folder in S3 or vice-versa:

```
  $ aws s3 sync myfolder s3://mybucket/myfolder --exclude *.tmp

    upload: myfolder/new.txt to s3://mybucket/myfolder/new.txt
```

The --exclude tag, as the name suggests, will exclude particular files and in this case, all files with the .tmp extension, and syncs everything else in the parent folder to the destination folder.

3.2 Cancer CellNet protocol

3.2.1 Cancer CellNet construction

Cancer CellNet is a derivative of CellNet and is designed to analyze the GRN status of cancer cell lines and tumors, by comparing them to the training data containing gene expression profiles of primary tumors. TtGRNs (tumor type specific gene regulatory networks) is defined as the steady state expression profile of a particular tumor type and we

measure the established ttGRNs for 12 types of primary tumor datasets, which serve as training data for Cancer CellNet. Based on the classifiers constructed as a result of training the program, expression profiles of query datasets, which in this case are cancer cell lines obtained from a particular tissue, can be used as inputs, to obtain the probability that the ttGRNs in the cell lines are indistinguishable from each tumor type in the training data set. For more information

All of the data used in this study is obtained from Affymetrix U133 Plus 2.0 microarray platform, and this was kept consistent across both training and query data sets in order to make the expression data comparable and to eliminate differences between platforms. The number of samples used to build the classifiers is tabulated below in Table 3.1. The raw expression files associated with these samples were curated from GEO (Gene Expression Omnibus) by our collaborator, Edroaldo Lummertz da Rocha and stored in S3.

Primary Tumor type	Number of samples
Acute Myeloid Leukemia	1013
Brain	423
Breast	2111
Colon	1559
Kidney	589
Liver	279
Lung	878
Melanoma	357
Ovary	928
Pancreas	178
Prostate	299
T-cell Acute Lymphoblastic Leukemia	814
Total	9428

Table 3.1 – Cancer CellNet training data sample set

3.2.2 Outline of the procedure

The overall procedure for Cancer CellNet, including the acquiring of data and preprocessing, training Cancer CellNet and running it on the query data set is depicted in

Figure 3.1. The procedure consists of two phases, Training Cancer CellNet where the tumor type classifiers are constructed based on the primary tumor data shown above is fed into the pipeline, and Querying Cancer CellNet, where the query data comprising of CCLE cell lines is fed as input to provide classification scores with respect to each tumor type classifier. The entire procedure is split into 10 subsections, each of which is explained in detail in the following sections:

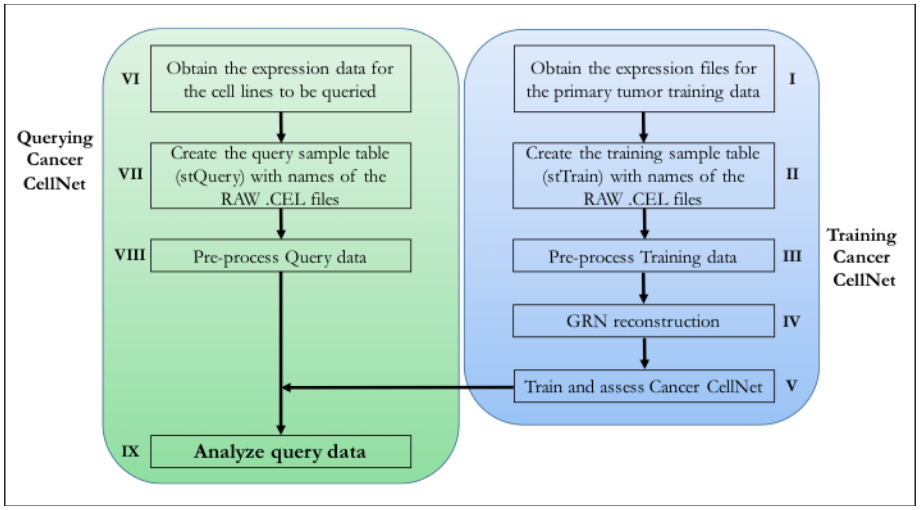


Figure 3.1 – Outline of the Cancer CellNet procedure

3.3 Training Cancer CellNet

Cancer CellNet (CCN) is an adaptation of the CellNet pipeline for cancer data. In order to train CCN, the raw .CEL files which contain expression data from the primary tumor samples listed in Table 3.1 were used. The detailed procedure for in this process is described below and the numbers in the titles correspond to the steps depicted in the above figure:

I. Obtain expression files for training CCN (by Edroaldo)

Manual curation of primary tumor data was done, by browsing through GEO manually for datasets corresponding to the twelve tumor types listed above, and care was taken to ensure the platform was consistent across the files. The compressed .zip files uploaded onto the S3

server into the designated folder and unzipped to get .CEL.gz files. This process was performed for all of the samples shown in Table 3.1 and will serve as input to the CCN pipeline, in order to train the classifiers.

II. Creating the training sample table

Using a spreadsheet editor, a csv (comma-separated value) file was created describing the annotation information for the training data samples. In this project, the training data sample table is called "stTrain.csv". The sample table is loaded into S3, so it can be called into the instance, during CellNet training. A subset of the training data sample table is shown in Table 3.2 below.

exp_id	sample_id	sample_name	cfname	description1	description2
GSE10245	GSM258551	GSM258551	GSM258551.CEL.gz	lung	endoderm
GSE10245	GSM258552	GSM258552	GSM258552.CEL.gz	lung	endoderm
GSE10282	GSM259617	GSM259617	GSM259617.CEL.gz	melanoma	ectoderm
GSE10609	GSM267321	GSM267321	GSM267321.CEL.gz	T-ALL	blood
GSE10971	GSM277776	GSM277776	GSM277776.CEL.gz	ovary	non-blood

Table 3.2 – Subset of the training sample table

Legend:

exp_id – The GEO experimental ID which identifies the parent dataset

sample_id – The specific GEO accession number for each sample that is a part of exp_id

sample_name – The name of the tumor sample assigned by the authors

cfname – The name of the raw expression data file obtained after unzipping the .zip file

description1 – The broad definition of the tumor type representing each sample

description2 – The germ layer from which the tissue originates from, during development

If sample_name is not specified by the authors of the dataset, it is initialized as the sample_id. The cfname serves as the link for the sample table to the raw data files stored in

S3, and helps in fetching the data during the execution of the program. Description1 will be the annotation for the classifiers to be built and hence, need to be consistent in annotation between all samples isolated from a particular tissue, within the sample table. Description2 helps differentiate between germ layers of origin and this becomes useful when the transcription factors that are part of the GRN are analyzed. TFs can play differing roles in various tissues and also during different stages of development and it becomes essential to have information about the germ layer of origin for the tumor tissue type, in order to get a better idea about the TFs that are dysregulated in these different tumor types.

III. Pre-processing training data

Pre-processing involves fetching the raw expression files, unpacking them, getting the raw data and normalizing them. In order to perform all these functions across the training data set, EC2 was used. After logging in to the AWS console, the CellNet-RNASEQ AMI was launched with c3.8x large instance type. Once the instance has initialized, terminal was launched and secure shelled into the running instance as described above. The steps involved in pre-processing are described in detail in the following points.

a) Screen is launched using the command "screen" typed into the terminal. This allows the processes to be detached and reattached, and is useful when the connection to the instance breaks or to toggle between multiple processes that are run in parallel in the same instance.

b) Create a new folder within the ephemeral drives, that are a part of the instance opened. This becomes necessary due to the large size of raw data that is about to be processed during training.

```
sudo mkdir /media/ephemeral0/data
sudo chown ec2-user /media/ephemeral0/data
cd /media/ephemeral0/data
```

c) The latest version of CellNet was installed and launched using the following commands

```
sudo R
library(devtools)
install_github("pcahan1/CellNet",ref="rpackage")
q()
```

d) R is launched within the same folder and the necessary packages and source codes are installed as shown below:

```
R
library(CellNet)
library(GEOquery)
source("cellnetr_utils.R")
```

e) The sample table created is loaded from S3, into the instance using the AWS CLI described above. That is initialized to the variable "stTrain", and this notation will be used consistently to denote the training sample table. The row names of the sample table need to be initialized to the sample id. This step is crucial in order to be able to link the annotation table and the expression matrix that will be generated at the end of the preprocessing step.

```
samptab <- read.csv("sample table training.csv")
row.names(samptab) <- samptab$sample_id
```

f) Individual functions are written to take as input, the cfname of the sample, and fetch the file from the S3 location specified, unzip them and introduce a new element in the sample table stripping the ".gz" from the name of the file, under the column titled "file_name".

g) Since the number of files that are present in the training data table is huge, all files cannot be unzipped and processed at the same time as it may crash the instance, and so, subsets of files need to be pulled up at a time. In order to group the training data into groups, their experimental id was used. All unique experimental ids were loaded into a vector.

```
sids <- unique(as.vector(samptab$exp_id))
```

h) All rows of sampTab containing samples belonging to a particular exp_id are grouped into a separate sample table, stTest, that will be used to obtain a normalized expression matrix:

```
stTest<-sampTab[which(sampTab$description1==sid),]
write.csv(stTest, file="stTrain_expn.csv")
stTrain <- expr_readSampTab("stTrain_expn.csv")
expTest <- Norm_cleanPropRaw(stTrain, "hgu133plus2")
```

where sid is an element in the vector sids. This process iterates through all values of sids to obtain expTrain for all samples in stTrain. Each expTest contains all the genes corresponding to the probes on the microarray platform, listed along the rows and each column represents a particular primary tumor sample. This normalization is a crucial step in preprocessing and all the expTest tables are given unique names and stored in S3 for other analyses.

i) Once the expression scores are obtained, the next step is to get a merged expression file and this can be achieved by combining all of the individual expTest tables as follows.

```
expTrain <- cbind(expTrain, expTest)
```

This command is executed at the end of every expTest generation step, prior to the start of the next iteration, so at the end of the loop, expTrain is a large expression matrix containing all the samples along the columns and all the genes along the rows. This R object is saved in S3.

IV. GRN construction

This step constructs tumor type specific Gene Regulatory Networks (ttGRNs), using the following function.

```
grnProp <- cn_make_grn(stAll, expAll, species = "Hs",
```

```
tfs = hsTFs, dLevel = "description1", dLevelGK =
"description2")
```

The arguments of cn_make_grn take as input, the sample table and the normalized expression matrix, created in the steps above. They also specify the species that the samples belong to, as the set of TFs differ across species. dLevel specifies the various tumor types we want ttGRNs to be constructed for and dLevelGK helps in distinguishing germ layer differences between tissue types, as the same TF may play different roles in different tissues and knowing the germ layer of origin for a particular tissue helps in constructing the GRN. The grnProp obtained as a result of GRN construction is also saved in S3 for future use, so the process does not have to be repeated.

The stringency parameters for constructing GRNs was adjusted that each ttGRN contained a minimum of 12 TFs that are associated with it. This becomes important to obtain

V. Train and assess Cancer CellNet

In this step, the classifiers formed as a result of the GRN construction are assessed. The way this is done is to split the training data into two parts (in this case, it was split into equal parts), and using one part to train Cancer CellNet and the other part serves as validation data. The results are stored in a list, that can be used to construct heat maps and PR curves, assessing the classifiers.

```
classifierPerformance <- cn_splitMakeAssess(stTrain,
expTrain, grnProp)
```

classifierPerformance is a large list that contains three parameters: PR curves, classifiers and the classRes, which is a large matrix containing the probability scores for the validation data. Good classifiers should result in high precision for high recall values. They are described in more detail in Chapter 4. The classification heatmap should result in all tumor

samples derived from a particular tissue classifying correctly as that particular tumor type. The code used to obtain these graphs is described in section 3.5.

Once the quality of the classifiers is assessed and deemed satisfactory, a large R object is created that contains all the information contained in the classifiers like normalized expression values (expTrain), the sample table for the training data (stTrain), the ttGRNs etc. This will be the object used while querying data in the steps below. In order to create this object, the following lines of code are used.

```
cnProc <- cn_make_processor(expTrain, stTrain, grnProp)
```

This is the most important object that is obtained as a result of training CellNet and this is also saved in S3 for use in the following steps.

3.4 Querying Cancer CellNet using Cancer Cell lines

In order to assess the fidelity of cancer cell lines that are in use in research, the CCLE database was used and the following steps were performed to feed the expression data from each of these cell lines into Cancer CellNet that is trained with primary tumor data.

VI. Obtained expression files for cancer cell lines

The compressed file containing the expression data from all 917 cancer cell lines was obtained, uploaded into S3 and unzipped to obtain all the raw .CEL.gz files in the required folder.

VII. Creating the query sample table

As described in part II, a csv file was created using a spreadsheet editor, a sample table was created, describing the annotation information for all the cell lines in the CCLE database, similar to Table 3.2 but with an extra column containing the names of the cell lines (cell_name)

was added. This sample table was named "stQuery.csv" and was uploaded into S3 so as to be called into the instance, while running Cancer CellNet.

S.No	Tumor Type	Number Of Cell Lines
1	AML	26
2	Autonomic Ganglia	20
3	Ball	16
4	Biliary Tract	7
5	Blood	109
6	Bone	25
7	Brain	46
8	Breast	57
9	Colon	55
10	Endometrium	25
11	Kidney	20
12	Liver	25
13	Lung	164

S.No	Tumor Type	Number Of Cell Lines
14	Melanoma	58
15	Oesophagus	26
16	Ovary	44
17	Pancreas	43
18	Pleura	9
19	Prostate	8
20	Salivary Gland	2
21	Soft Tissue	16
22	Stomach	34
23	T-ALL	15
24	Thyroid	11
25	Upper Aerodigestive Tract	31
26	Urinary Tract	21

Table 3.3 – Cancer CellNet query data sample set – CCLE cancer cell lines

The cancer cell lines in CCLE were derived from 26 major tissue types and they are listed across two columns in Table 3.3. The ones highlighted in green are those tumor types that are a part of the training data set.

VIII. Pre-processing query data

Similar to section III above, the raw data for the cancer cell lines is extracted and the expression data is normalized. Since the aim of the project is to assess cancer cell lines based on their fidelity with respect to primary tumors, it makes sense to query all cell lines belonging to a particular tissue type together, as opposed to querying all cell lines together. The latter will result in a very populated heatmap that becomes hard to interpret.

The following steps are performed to obtain expQuery, which is a matrix containing normalized expression data for all cell lines belonging to a particular tumor type and is repeated for all distinct tumor types listed in Table 3.3.

a) Launch screen by typing "screen" in the terminal. This enables us to log back in, when the instance pipe gets broken.

b) Create a folder within the ephemeral drive as shown above and download the sample table and the cnProc from S3. The cnProc created at the end of training and the query sample table are downloaded into the working directory inside the instance. These are initialized to variables cnProc and sampTab respectively.

```
source("cellnetr_utils.R")
cnProc <- utils_loadObject("cnProc.rda")
sampTab <- read.csv("sample table query.csv")
row.names(sampTab) <- sampTab$sample_id
```

c) Since we are grouping the cell lines as per their tissue of origin, the sids vector will be initialized as shown below, and not by experimental ID.

```
sids <- unique(as.vector(sampTab$description1))
```

The following steps will be looped for every element in the vector "sids". This way, all cell lines belonging to every tissue type gets extracted and their expression data analyzed.

d) A subset of the query sample table is taken from sampTab, the raw files fetched from S3, unpacked, and the expression data normalized as described in step (h) in section 3.3. The individual expression matrices are combined into a large expression matrix, that contains genes along the rows and the cell lines along the columns. This normalized expression matrix is called "expQuery".

IX. Analyze query data

Once we have the expQuery, and the stQuery with rownames equal to the sample_id, and the cnProc obtained as a result of training CellNet, we are ready to query the data against the classifiers established by the training data set. The following command runs CellNet and compares the expression data of each sample against the training classifiers:

```
tmpAns <- cn_apply(expQuery, stQuery, cnProc)
```

The tmpAns is an R object of class cnRes and it can be obtained for each of the 26 tissue types, and saved in S3 with the file name containing the tissue type it represents. This way, a particular subset out of all the tissue types can be chosen for further analysis. Using this result, the best cell line models for each type of tumor can be identified.

3.5 Additional function definitions

3.5.1 Creating classification heat maps

In order to create heat maps, the "pheatmap" function was used. As arguments for this function, the matrix that contains the classification probability scores for each specific tumor type, and color cues to specify the range of values were fed. The following code does this for one specific example:

```
newHm <- function(cnRes){
classMat <- cnRes$classRes
cools <- colorRampPalette(c("black", "green", "yellow"))(100)
pheatmap(classMat, col = cools,
         breaks = seq(from = 0, to = 1,length.out = 100),
         border_color = bcol, cluster_rows = FALSE,
         cluster_cols = TRUE,
         fontsize = 5)
}
library(pheatmap)
tmpAns <- utils_loadObject("tmpAns_ovary.rda")
newHm(tmpAns)
```

3.5.2 Calculating Network Influence Scores (NIS)

Network Influence Score indicates the extent of dysregulation of the transcriptional regulator and its target genes in the sample, weighted by the importance of that TF to that particular tumor type GRN. This can be used to prioritize hypotheses to explain the molecular differences between cell lines and tumors. It is defined using the following formula:

$$NIS(TR) = \sum_{i=1}^{n} (Zscore(target)_{TT} * weight_{target}) + n * Zscore(TR)_{TT} * weight_{TR}$$

where TR is a transcriptional regulator and n is the number of genes in the GRN for a particular tumor type (TT). Weight is the difference in expression with the mean expression of the gene in that tumor type, calculated in the classifiers. The weight of the TR and its targets is calculated by CellNet by their mean expression in the tumor type samples that are a part of the training dataset when searching for inappropriately silenced factors.

The code used to calculate NIS for each cell lines grouped per tumor type, takes as input the cnProc that is obtained as a result of training CCN and the cnRes obtained for that particular tumor type cell lines at the end of running Cancer CellNet. The syntax is as follows:

```
tfScores <- cn_nis_mod(cnRes, cnProc, subnet, tumor-type)
```

where subnet refers to a subset that is taken for evaluation, like all "prostate" samples for example and tumor-type refers to the name of the classifier, whose ttGRN is being compared against, which in this case can be any one of the 12 tumor types used in training CellNet. The function cn_nis_mod calculates the NIS scores for all TFs that are a part of the ttGRN for that tumor type. The expanded code for cn_nis_mod can be found in github.

The variable "tfScores" contains the NIS for each of the TFs associated with the tumor type specific GRN. These values can be negative or positive depending on increase or decrease in the level of expression of the TF, as compared to the corresponding value in the tumor type classifier. In order to consider the fold change and not the direction, absolute values of NIS is obtained and TFs are arranged in decreasing order of this fold change for each cell line, with the most misregulated TF being on top. The frequency of occurrence of a particular TF among the top 10% in each cell line is calculated and tabulated, and this process is repeated for every tumor type analyzed.

3.5.3 Obtaining GRN status

GRN status is a measure of the extent of establishment of a specific ttGRN in the query set. This is defined as the weighted sum of the z scores of the genes in the GRN. The formula used to calculate raw GRN status is as follows:

$$RGS(x)= \sum_{i=1}^{n} (Zscore(gene)_{TT}*gene\ weight_{GRN})$$

where n is the number of genes in the GRN and gene weight is the proportional expression in the tumor type. The raw GRN status scores are normalized to the final GRN status that gets plotted, in order for a better score to indicate a better resemblance to the training data GRN.

$$GS(x)= 1000-\frac{RGS(x)}{(\sum_{j}^{k} RGS(y))\ /k}$$

where the denominator term represents the average raw GS of tumor type samples in the training data. This is plotted by providing the function with the cnProc used to construct the classifiers and the cnRes object obtained at the end of CellNet querying, and mentioning the tumor type to be compared against. The syntax is as follows:

```
cn_barplot_grnSing(cnResQuery, cnProc, GRN_tumor_type,
                   training_GRN , query)
```

where GRN_tumor_type is the tumor type specific GRN, to which the extent of establishment is measured in the query samples. training_GRN is the extent of the GRN establishment for cell_type_1, in the training dataset. This serves as a control relative to which the query samples are measured. query is a string containing the order of query samples to be displayed in the final plot.

3.5.4 Obtaining citations from Pubmed

In order to obtain the number of citations in Pubmed for each of the cancer cell lines in our study, published in the context of cancer, the search term included the terms "AND cancer" along with the name of each cell line, to limit the hits to only those publications that were in the context of this project. The following lines of code was used to automate this:

```
tumor_type <- read.csv("sample_table_tt.csv")
cell_lines <- as.data.frame(tumor_type$cell_line)
for (i in 1:nrow(cell_lines)){
  cell_lines[i,2] <- as.character(paste(query_aml[i,1], "AND
cancer*[tw]"))
}
query <- as.vector(query_aml[,2])
names(query) <- query_aml[,1]
df <- PubMedTrend(query)
```

where the sample table for a particular tumor type, containing all the names of cell lines that are derived from that tumor is extracted and the term "AND cancer" is appended to the names. This is then fed to the function PubMedTrend that has been expanded in Github. The output of this function results in a data frame containing the number of hits for the search-term in Pubmed, listed year-wise. That dataframe can be written into a csv file and used for further analysis.

Chapter 4 - Results

4.1 Training Cancer CellNet and assessing the classifiers

As described in section 3.3, the CellNet pipeline has to be trained with primary tumor data in order to construct the ttGRNs for all primary tumor types, in order to be able to classify cancer cell lines based on their classification scores obtained after passing them through CCN. Cancer CellNet was trained using data obtained from samples listed in Figure 3.1 (initial training of CCN performed by Kathleen diNapoli). The ttGRNs constructed at the end of this training was analyzed for checking the quality of training data using two tools:

a. Classification heatmap

The classifier performance R object obtained at the end of training, as described in step V of the CellNet procedure, contains the classification scores for those samples that were a part of the primary tumor expression data set, that was used as query to assess the ttGRNs constructed at the end of CellNet training (Split and assess function described above). These classification scores are represented in the form of a heatmap as shown in Figure 4.1.

The rows represent the 12 primary tumor types that constitute the training data set and the columns are arranged in order of the annotated tumor type of that sample. For example, the first 507 columns are samples that are annotated in the training sample table as being derived from AML, and hence, should show a perfect classification with AML as the ttGRNs for this tumor type was constructed using this data as input. And this logic applies for all of the tumor types that are a part of the training data set. If there are certain samples that do not classify to the right tumor type, that indicates a bad tumor dataset.

In order for the training to be successful, and the ttGRNs be considered valid representations of the actual tumor type, the heatmap should show little to no such misclassification, and those sample that classify correctly should also exhibit a classification score close to 1. This becomes crucial because if there is a significant portion of the training data that is either mis-annotated to be from a tumor type that is different to that it is annotated to be from, or if the expression data is not of good quality, it will affect the ttGRN construction for that tumor type. This may result in errors while using these parameters to query the cell line expression data, resulting in either false negatives or false positives.

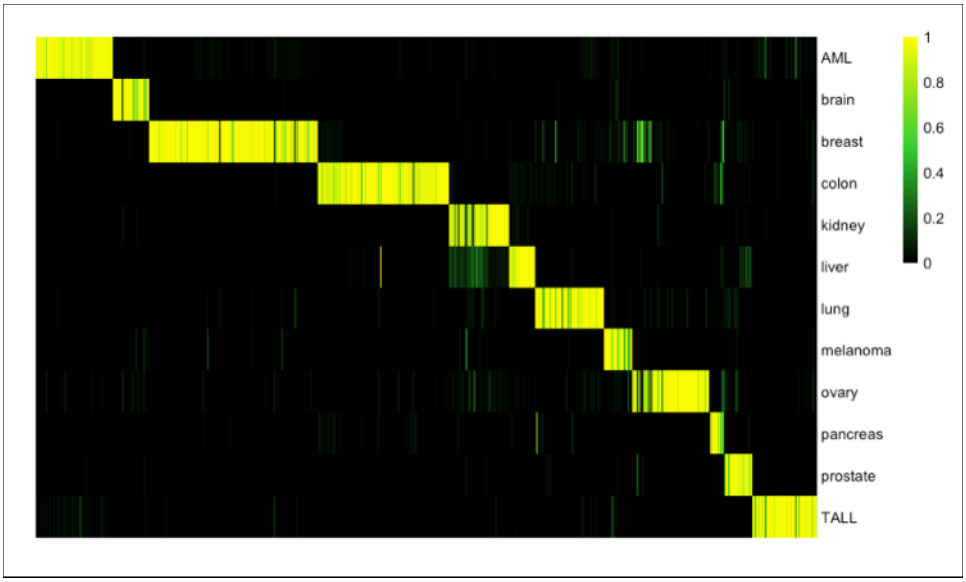


Figure 4.1 – Classification scores to test the quality of training data

Analysis of Figure 4.1 shows good classification for samples derived from most tumor types, except kidney, melanoma and ovary. A significant number of kidney tumor samples show an overlap with liver, while melanoma and ovary are showing a non-specific classification with breast. However, with these few exceptions, most of the validation dataset classified in accordance with their tumor types of origin, with high probability scores. Overall, this training of CellNet is considered successful.

b. Precision-Recall curves

This analysis depicts both the sensitivity and precision of the classifiers to accurately classify a sample correctly into its tumor types. Precision and recall are defined using the following formulae:

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

In words, precision translates to the ability of the system to detect true positives among all "perceived" positives and recall is detection of true positives from "actual" positives. These terms have an inverse correlation as the more precise it gets, the lesser its recall is going to be. The PR curves represent the relationship between precision and recall, and ideally, the precision should remain high for almost all values of recall.

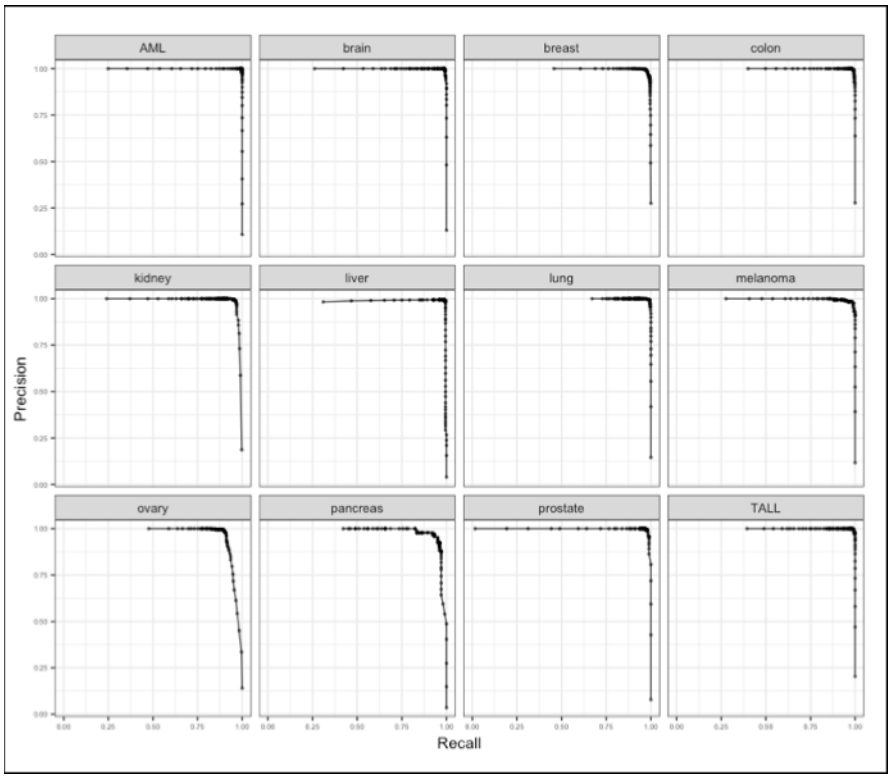


Figure 4.2 – Precision-Recall curves for CellNet generated tumor type classifiers

PR curves were plotted for each of the tumor type classifiers and represented in Figure 4.2. These were obtained from the classifier performance R object that is obtained at the end of step V in the CellNet pipeline (Refer Chapter 3). From the PR curves obtained at the end of training, all the tumor type classifiers show very high precision values with respect to high recall. Some tissues like ovary and prostate show lesser precision than others, but they are 100% precise up until recall values of 0.9, which indicates a very good classifier.

Thus, training of Cancer CellNet using the primary tumor data was successful in establishing ttGRNs and was used for querying cell line expression datasets, described below.

4.2 Querying cell lines from tissue types in the training dataset

The dataset used to train CellNet and build the classifiers is a compilation of expression data derived from primary tumor samples originating from 12 tissue types. Cell lines derived from these tumor types will thus serve as the first set of query datasets, in order to assess the validity of these cell lines as accurate models of the corresponding tumors.

4.2.1 Acute Myeloid Leukemia (AML) cell lines

Acute Myeloid Leukemia or AML is a tumor of the hematopoietic and lymphoid tissue and is characterized by proliferative, abnormally differentiated or de-differentiated cells of the hematopoietic system infiltrating the bone marrow, blood and other tissues[39]. The genome of AML tumors show relatively lower mutation profiles as compared to other types of cancers[40], with an average of 13 mutated genes[41], according to the Cancer Genome Atlas Research Network.

A number of cell lines have been established from AML patients and the CCLE[12] profiles 26 such cell lines. Running the expression data obtained from these cell lines through CellNet resulted in classification of all cell lines as AML-derived, as shown in Figure 4.3. The

figure shows all 26 cell lines along the x-axis and the heat map represents the probability of the query exhibiting GRN genes to an extent that is indistinguishable from the tumor type mentioned along the y-axis. The AML cell lines are ordered in increasing order of their probability scores for AML tumor type, the lowest scoring line being on the left.

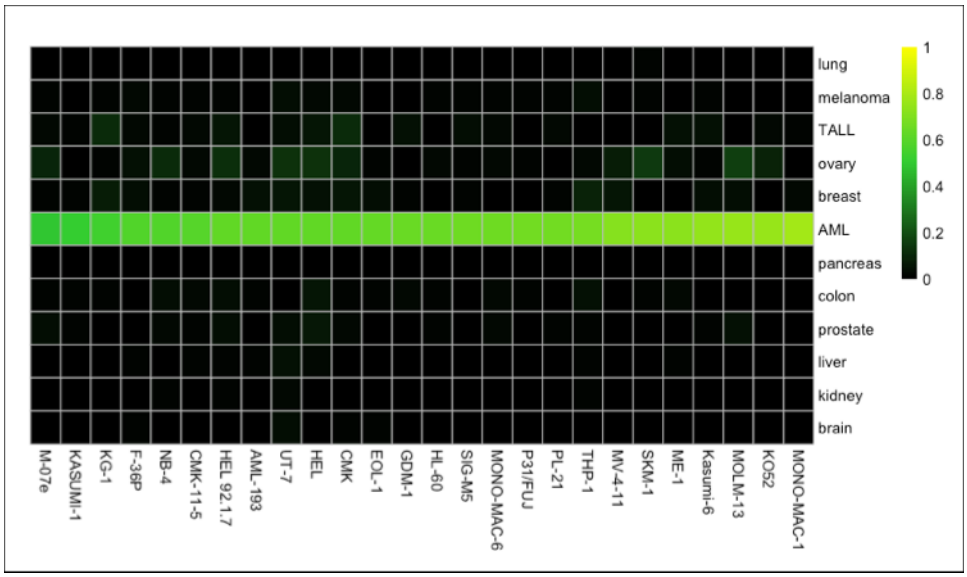


Figure 4.3 – Classification of AML cancer cell lines

AML cell lines are derived from different cell types; for example, HL-60 is a pre-myeloblast cell line whereas AML-193 is a monocyte derived[42]. Given this diversity, the robustness of the classification shows that the underlying GRNs are similar across all AML cell lines and they also serve as a true representation of primary AML tumors, since the probability of classification for all cell lines was higher than 50% and show almost no non-specific classification with other tissue types. From this classification heatmap, it can be inferred that there are some cell lines which show minor probability of classifying to other tumor types, like the ovary. In order to better understand why this occurs, the tumor-type specific GRNs (ttGRNs) for each of the cell lines that can be extracted from the result of running Cancer CellNet (cnRes object), were analyzed ("tmpAns" – refer step IX of the CellNet protocol in Chapter 3).

There were 71 component transcription factors (TFs) in the AML-specific-GRN that were identified as a result of training Cancer CellNet with AML primary tumor samples. The Network Influence Score (NIS) for each of these TFs was calculated across all 26 AML-derived cell lines, as described in the methods section. The frequency of the most dysregulated TFs across all queried cell lines are tabulated in Table 4.1.

Transcription Factors	Frequency (%)
ELF1	100
NFIL3	100
RNF10	100
MNDA	73.08
ZNF394	61.54
LYL1	50
HHEX	46.15
CEBPA	42.31

Table 4.1 – Frequency of maximally-misregulated ttGRN specific TFs across AML-cell lines

ELF-1 is E-74 like ETS Transcription Factor 1 and it encodes a protein that is primarily expressed in lymphoid cells and is known to act as both activator and inhibitor of gene expression[43]. Importantly, it plays a crucial role in the transcription of the TCR ζ chain and expression of this protein is reduced in many hematological malignancies including AML. Expression level of ELF-1 was measured from 33 patients suffering from AML and it was observed that the protein was overexpressed in all of them, as compared to healthy controls. It is believed that the down-regulation of the TCR ζ chain in AML leads to a feedback regulation in the expression of ELF-1[44]. However, in all of the AML derived cell lines that show relatively poorer classification scores with the AML classifier, the ELF-1 expression is ~2 fold lower than the mean expression level in the ttGRN.

NFIL3 (Nuclear Factor, Interleukin 3 Related) is a transcriptional regulator that binds DNA as a homodimer to ATF (Activation Transcription Factor) sites and represses transcription of those genes[45]. Thus, it acts as an inhibitor of the circadian rhythm genes PER1 and PER2[46] but activates transcription of interleukin-3 promoter in T-cells. Since it restricts FOXO binding onto the promoter elements, this protein is referred to as a "survival factor", and is up-regulated in AML "stem-cells". In almost all of the cell line models, this gene's expression levels are about one fold lower than primary tumor cells, but does not vary significantly between cell lines.

The third TF that is dysregulated in all of the AML cell lines is RNF10 (Ring Finger Protein 10), which has a ring finger motif that facilitates protein-protein interactions[47]. Interestingly, the specific function of this protein still remains unknown and its possible relationship with AML is unexplored.

4.2.2 T-cell Acute Lymphoblastic Leukemia (T-ALL) cell lines

T-cell acute lymphoblastic leukemia is an immature, aggressive hematological tumor derived from early T-cell progenitors. Malignant hematopoietic cells expressing immature markers for the T-cell infiltrate the bone marrow resulting in increased white blood cell numbers and hematopoietic failure[48]. Based on gene expression signatures, there are multiple unique biological subgroups of T-ALL clinically, which result in different cellular phenotypes[49].

T-ALL is considered a heterogeneous cancer type based on its immunological phenotype. Burger et al in 1999[50], analyzed the differences between all T-ALL derived cell lines used in research in a comprehensive manner. They performed southern blotting, immunophenotyping and TCR protein expression studies and their results showed a high

degree of heterogeneity among the 16 cell lines they analyzed. However, there was no comparison between the cell lines and primary tumors in this study.

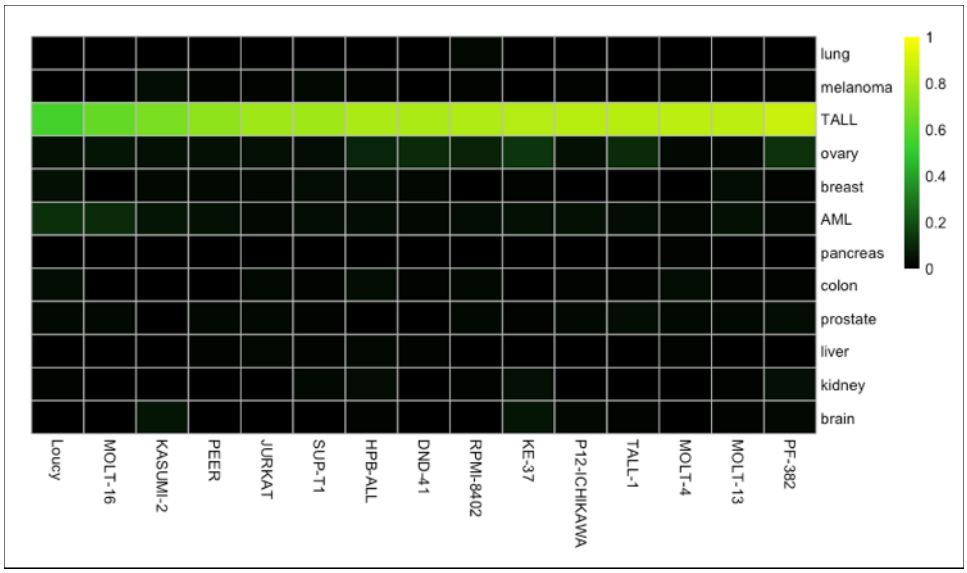


Figure 4.4 – Classification of T-ALL cancer cell lines

Fifteen T-ALL cell lines were characterized in the CCLE database, and analyzing the GRN-based tissue classification of these cell lines resulted in highly specific classification into the T-ALL tumor type, as shown in Figure 4.4. Thus, it can be concluded that although among T-ALL cell lines and in-vivo tumors, there exists phenotypic heterogeneity, the cell lines that are used to research this type of leukemia recapitulate the transcriptomic characteristics of the primary tumors with fidelity, few cell lines with higher similarity to the tumor profile than others.

Among the cell lines, the one showing maximum classification with the T-ALL tumors was PF-382, with a score of 88% and it was isolated from a 6 year old female with a second relapse of the disease[51]. In order to understand the molecular players that are different in cell lines as compared to the primary tumors, the TFs that are a part of the ttGRN profile are analyzed and those that deviated from the tumor expression level in 50% or more of the cell lines are tabulated in Table 4.2 below.

Transcription Factors	Frequency (%)
JUND	86.67
TNFAIP3	86.67
TCF7	80
LEF1	73.33
CHCHD2	66.67
MTF2	66.67

Table 4.2 – Frequency of maximally-misregulated ttGRN specific TFs across T-ALL cell lines

JUND codes for the JunD protein that is a functional component of the AP-1 transcription factor complex and its major role in normal cells is to prevent p-53 dependent senescence and apoptosis[52]. It is a proto-oncogene and is commonly associated with diseases like Leukemia, especially adult T-cell leukemia[53]. There are no known reports linking the activity of this gene to be responsible in T-ALL till date. However, in a T-ALL cell line called CCRF-CEM, which was not a part of this study, p53 induced apoptosis was found to be enhanced[54] and in the cell lines that are a part of our analysis, the expression levels of JUND appears to be down-regulated. If the trend continued across all T-ALL derived cell lines, this may result in increased incidence of apoptosis in cells during culture.

TNFAIP3 is Tumor Necrosis Factor Alpha-induced protein 3 and it is a ubiquitin-editing enzyme. This protein is actively involved in inflammatory and immune pathways in the body. This gene, along with TCF7 (T Cell Factor 7), which is also a part of Table 4.2, are responsible for the regulation of T-cell receptor expression on the surface of T-cells. Although T cell surface markers are aberrantly expressed in T-ALL disease manifestation, the down-regulation of these genes in culture results in deviation from cell surface marker profile of in vivo cancer cells.

4.2.3 Central Nervous System (CNS) tumor derived cell lines

There are three major classes of brain tumors namely astrocytomas, oligodendrogliomas and meningiomas[55]. Table 4.3 shows the number of samples in both the training and the query dataset, derived from specific types of brain tumors. It can be seen that ~62% of the training data comprises of glioblastoma tumors, which are the most commonly occurring subtype of glioma in humans. Glioblastomas are also referred to as "Grade 4 Astrocytomas" and are the most malignant form of brain tumors derived from the supporting cells of the CNS, namely the glia[56]. Thus, it is also the most researched among other forms of gliomas and almost 70% of all CNS cancer derived cell lines originate from glioblastomas.

In 2007, the World Health Organization (WHO) classified central nervous system tumors according to the microscopic similarities between putative cells of origin, predominantly done by Hematoxylin and Eosin (H&E) staining of tissue sections, and their presumed level of differentiation[57]. This eliminated the diversity between tumors derived from the same class of cells, for example, all astrocytic tumors were classified together irrespective of their diversity in clinically manifested phenotypes. However, the latest WHO classification, published in 2016[58], updates this classification and incorporates the various differences between tumor types and provides an updated synopsis of tumor classification of the CNS.

Primary tumor type	Number of samples in training dataset
Astrocytoma, grade 2	7
Astrocytoma, grade 3	19
Ependymoma	46
Glioblastoma	204
Glioblastoma, grade 4	60
Medulloblastoma	22
Oligodendroglioma, grade 2	38
Oligodendroglioma, grade 3	12
Pilocytic Astrocytoma	15
Total	**423**

Primary tumor type	Number of cell lines in query dataset
Glioblastoma	33
Gliosarcoma	1
Non-specified	11
Oligodendroglioma	1
Total	**46**

Table 4.3 – Types of gliomas and composition of the training and query data sets

The CCLE database was compiled and published in 2012 and thus, used the 2007 broad classification to group cancer cell lines. Therefore, when the 46 CNS derived cancer cell lines were run through the CellNet algorithm to obtain tumor type classification, the results showed no strong correlation to the brain tumor training dataset, as shown in Figure 4.5. The heatmap has been arranged in increasing order of probability of the queried cell lines to classify as brain, and it can be seen that a majority of cell lines do not classify as brain or any other tumor type in general. This can also be explained by the fact that the CNS-tumors originate in cell types that are completely different from one another in their morphology, function and hence, their inherent gene expression profiles. But, grouping them under the broad umbrella of CNS-derived tumors eliminates this diversity, and while averaging out the primary tumor profiles to get a consensus ttGRN, these differences may have gotten filtered out.

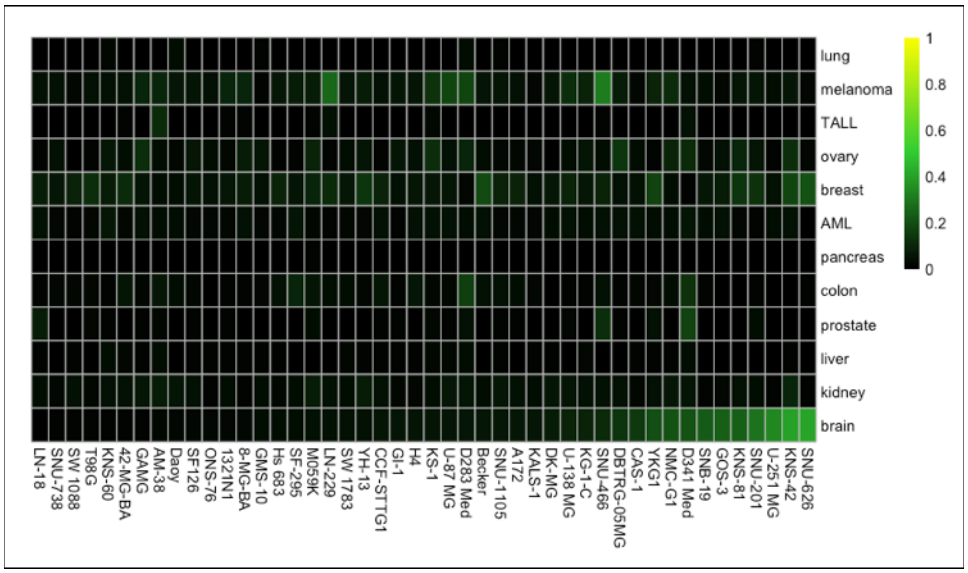


Figure 4.5 – Classification of CNS tumor derived cell lines

The maximum probability score for cell lines to classify as brain, obtained from Cancer CellNet is 0.448 and this shows only moderate specificity for these cell lines to primary CNS tumors. The low scoring cell lines are predominantly those that have a non-specific CNS tumor origin and are underrepresented in the brain tumor training dataset, as can be seen from

Table 4.3, although some glioblastoma cell lines are also present in this category. Overall, ~25% of the CNS derived cell lines show a tendency to classify as brain. The extent of brain GRN status achieved by CNS tumor derived cell lines is depicted in Figure 4.6, to better emphasize this point.

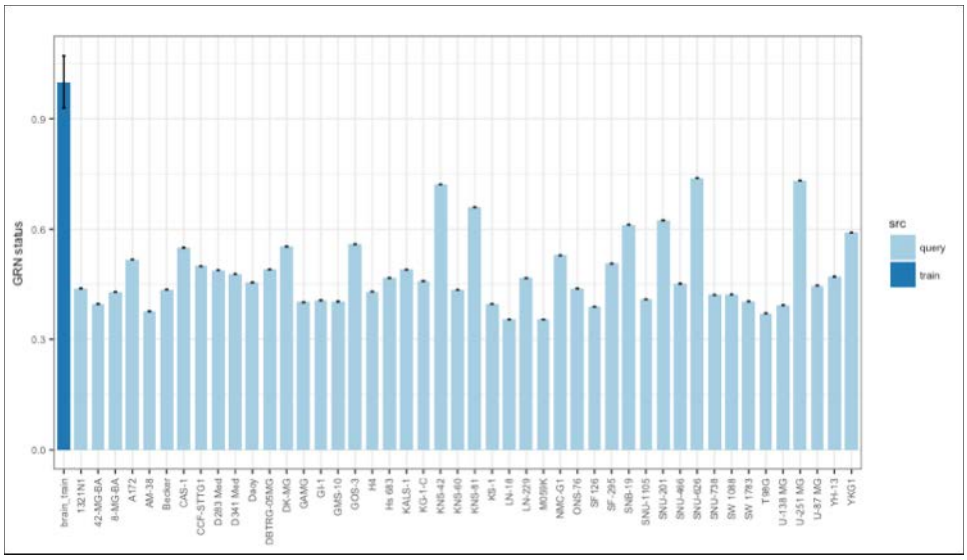


Figure 4.6 – Brain ttGRN status for CNS tumor derived cell lines

Among the cell lines profiled, the ones that classify the most as brain tissue and show a GRN status that resembles closest to the primary brain tumors are SNU-626, KNS-42 and U-251 MG, with classification scores of 0.41, 0.40 and 0.34 respectively. Analyzing the heatmap classification (Figure 4.5) for these three cell lines shows that SNU-626 and KNS-42 have non-specific classification with some probability, to other tissue types like breast and ovary, which is absent in U-251 MG. Thus, based on these results, the use of U-251MG as an in vitro model for CNS tumor studies is recommended.

4.2.4 Kidney tumor derived cell lines

Kidney tumors can be of various types like transitional cell carcinoma, sarcoma, Wilms tumor and the most common type of kidney tumor, which accounts for 85% of all diagnoses, Renal Cell Carcinoma (RCC). This cancer is usually treated with invasive surgery

to remove the tumor as it is resistant to chemotherapy and radiation[59]. RCC is characterized by highly vascular tumors, targeted therapies such as inhibitors of mammalian target of rapamycin (mTOR) and vascular endothelial growth factor (VEGF) pathways are used to treat metastatic renal carcinomas.

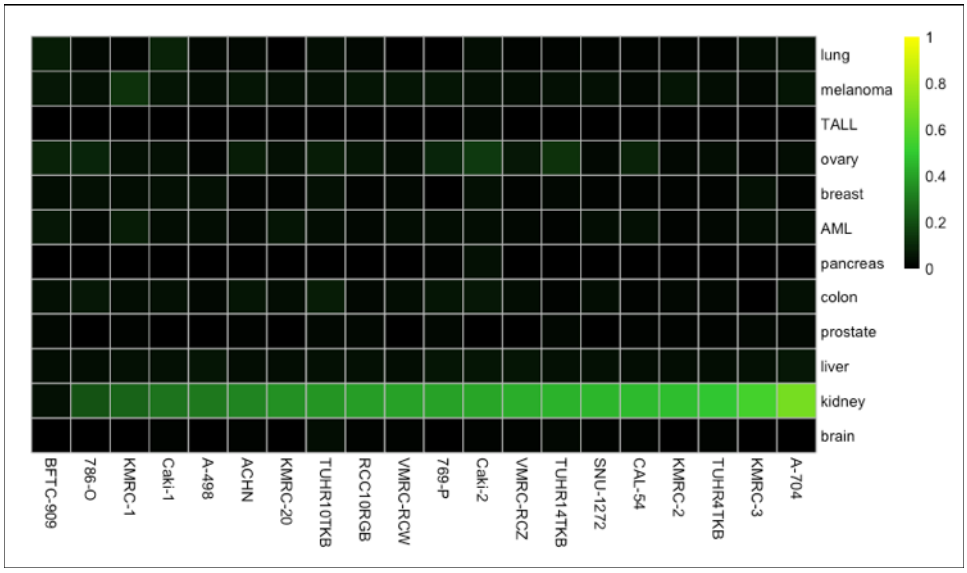


Figure 4.7 – Classification of renal cancer cell lines

CCLE contained 20 kidney cancer derived cell lines and 15 of those were from RCC and the tumor type of the other 5 were not specified in their original publications. Running their expression data through CellNet trained with primary tumor samples resulted in a heat map depicted in Figure 4.7. The cell lines in the heat map are arranged in increasing order of probability of being classified as a kidney derived cell line. 19 of these cell lines are rightly classified as kidney tissue-derived, except BFTC-909 which shows negligible probability with the kidney derived ttGRN classifiers.

BFTC-909 cell line was derived from a grade III transitional cell carcinoma from the bladder of a 64-year old male. Since it is derived from the bladder but annotated as a renal cell line, the absence of renal tumor specific GRNs would have resulted in a poor classification score for this cell line to be indistinguishable from primary kidney cancer. The A-704 cell line

shows highest resemblance to the kidney tissue classifier and there is no significant overlap with other tissue types and hence, can be considered to be the best model for renal cancers, among the cell lines profiled in this study.

Transcription Factors	Frequency (%)
NOSTRIN	100
CEBPD	75
ZNF395	65
HSF4	55
NR1H4	55
KLF9	20

Table 4.4 – Frequency of occurrence of dysregulated kidney ttGRN specific TFs across cell lines

In-depth analysis of the TFs that are dysregulated in the kidney cancer cell lines reveal that the NOSTRIN gene is down-regulated in all 20 cancer cell lines. NOSTRIN stands for Nitric Oxide Synthase Trafficking and this protein is expressed in highly vascular tissues like lung, kidney, heart and placenta[60]. Its function involves binding to eNOS (endothelial Nitric Oxide synthase), the enzyme that synthesizes nitric oxide within the cell and attenuates its function[61]. Thus, it has a significant role in processes like neurotransmission, vascular homeostasis and inflammatory responses. Knockdown of this gene has been shown to deter glomerulus function in zebrafish[62]. In human renal cancers, the expression of this protein is found to be higher than that of normal cells. Nitric Oxide dependent eNOS signaling has been shown to promote cancer cell survival under hypoxic conditions in many tumors[63] but since cell lines do not have the glomerulus architecture and are not cultured under hypoxic conditions, it is possible that the NOSTRIN gene is constitutively down-regulated in those clones that now make up these renal cell lines. All cancer cell lines in culture presently are believed to be clonal populations of fast-growing clones.

4.2.5 Ovarian cancer cell lines

Ovarian cancer victimizes more than 100,000 women globally, and is the most lethal gynecological malignancy in the United States[64]. Epithelial ovarian carcinomas are divided into 4 subgroups: serous, endometroid, clear cell and mucinous among which serous carcinomas are the most prevalent and aggressive. A particular subtype, high grade serous ovarian carcinoma (HGSOC), is responsible for ~66% of ovarian cancer fatalities[65] and is one of the most extensively studied ovarian cancers. Due to the high frequency of occurrence of the serous subtype, primary tumor databases like TCGA[66] are skewed with respect of number of studies of ovarian cancer subtypes like HGSOC and this point becomes an important one to note as one goes about analyzing the cell lines that are a part of this study.

Domcke et al[13], in 2013, performed a bioinformatics-based comparison of gene expression and mutational profiles of 47 ovarian cancer cell lines that are widely used in research, with that of primary ovarian tumor samples, obtained from the CCLE and the TCGA respectively. Using information on mutation frequency, copy number variations obtained from mRNA expression data, the study derived and assigned a 'suitability score' (refer Methods) for cell lines that best represent serous carcinomas obtained from patients, and ranked their cell lines with respect to this score. Their suitability score (S) was a combination of Pearson's correlation scores of cell lines with the average expression of tumors, factored with presence/absence of mutations in certain key genes, and hyper-mutated cell lines were also penalized. They concluded that cell lines most commonly used as in-vitro models do not resemble the parent tumors.

Here, we studied the properties of 44 cell lines that are a part of CCLE, and compared them with our extensively compiled primary tumor training data set. As can be seen from Figure 4.8, the cell lines are grouped in order of increasing probability of classifying into an

ovarian cancer-derived cell line. The cell lines corresponding to the highest probability measures also show reduced non-specific classification, as compared to those with low probability scores. JHOM-2B, the cell line that shows the least classification score with ovary but a reasonably high score with colon, had a significantly decreased expression of Pax8, a TF that is found in almost 90% of ovarian cancers[67] and expressed at basal levels in all other ovarian cancer cell lines. Another ovarian cancer oncogene, YAP1[68], showed reduced expression exclusively in this cell line, as compared to others and this would have led to reduced classification as an ovarian cancer.

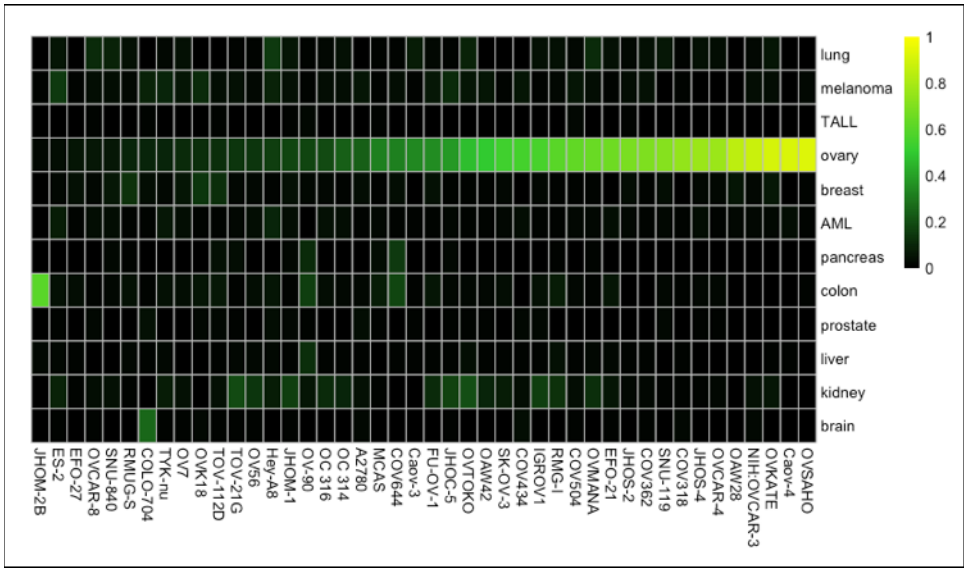


Figure 4.8 – Classification of ovarian cancer cell lines

Upon comparing the top 10 cell lines that most resemble ovarian cancer as a result of this study with the Domcke et al study, it is observed that there is 80% overlap between the top cell lines picked out by these studies. The two exceptions are KURAMOCHI and TYKnu. The cell line KURAMOCHI, which ranked highest in the 2013 study was excluded from our query dataset for Cancer CellNet, as its expression data was generated by the Affymetrix U133 array, which was not the platform our training data is generated from. Surprisingly, TYKnu, which appears in their list is featured very low in the cell line ranking score generated from

Cancer CellNet. It only demonstrates an 18% probability to resemble the training dataset.

The suitability score for all the cell lines obtained from their study and Cancer CellNet classification scores are compared in Figure 4.9. As can be seen, all of the cell lines our pipeline predicts to be good ovarian cancer models also exhibit a high suitability score. However, unlike the limited scope for resolution between the cell line models in the 2013 study, our analysis is able to provide a clear distinction between all the cell lines, as can be seen by their wide range of x-values in the graph shown below. Almost 50% of the Domcke et al high scoring cell lines performed poorly when their GRNs were compared with those of primary tumors.

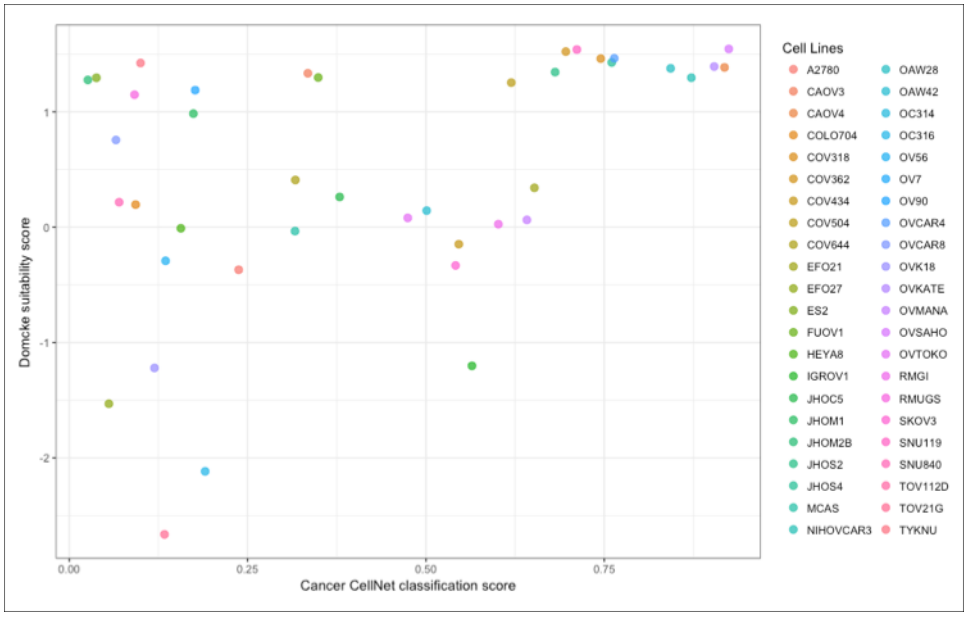


Figure 4.9 – Ovarian cancer cell line fidelity measurement comparison across two studies

The TFs that are deviating from the primary tumor expression profile were analyzed, in order to determine the molecular mechanisms in which tumor cells in culture differed from their in vivo counterparts. Surprisingly, only 3 TFs were found to significantly vary consistently across all cell lines, with respect to the tumor profiles and these were WT1, MEIS1 and SOX17. WT1 is Wilms' Tumor protein 1, an essential protein in the development of the urogenital system in humans[69] and as the name suggests, the absence or inactivation of this

protein leads to the formation of Wilms' tumor. It is present in almost 95% of all primary serous ovarian carcinomas[70,71], but strangely enough, it is consistently at-least two-fold down-regulated in almost all the ovarian cancer cell lines. However, the top 5 cancer cell lines models predicted by Cancer CellNet showed close to basal levels of expression for this protein.

4.2.6 Breast cancer cell lines

Breast cancer is the second most widely occurring malignancy in world population, and the most researched cancer in the United States, where $600 million was spent in the year 2016 alone. Due to significant progress made in this front the disease prognosis after an early diagnosis is very good[72] leading to an overall increase in life expectancy of breast cancer survivors (Figure 1.2). Breast cancer can be sporadic or familial, where the former is an occurrence due to the environment of the individual while the latter is genetically transmitted[73]. Traditionally, breast cancer has been grouped into subtypes based on the presence of hormone receptors (estrogen and progesterone receptors) on cells, pathology or grade of the cancer[74]. However, with advances made in molecular and gene expression profiling, this classification is being refined to better indicate prognosis and pathogenicity.

Cell line	Origin	Age (years)	Pathology
BT20	Breast	74	Invasive ductal carcinoma
MDA-MB-231	Pleural Effusion	51	Adenocarcinoma
MDA-MB-435	Pleural Effusion	31	Invasive ductal carcinoma
MDA-MB-468	Pleural Effusion	51	Adenocarcinoma
MCF-7	Pleural Effusion	69	Invasive ductal carcinoma
SkBr3	Pleural Effusion	43	Adenocarcinoma
T47D	Pleural Effusion	54	Invasive ductal carcinoma
ZR75.1	Ascites	47	Invasive ductal carcinoma

Table 4.5 – Origins of the most commonly used breast cancer cell lines[75]

The first breast cancer cell line was established in 1958 and named BT-20. The most commonly used breast cancer cell line in research is MCF-7 as it is ideal for studying cancers

with hormone receptors as it expresses the estrogen receptor and hence, serves as a perfect model to study hormonal response. This is the most cited breast cancer cell line and was derived from pleural effusions in 1978. Most of the commonly used breast cell lines are derived from pleural effusions or aspirates and hence, are derived from tumor metastases and not the primary tumor itself. A list of the most cited breast cancer cell lines with their source is tabulated in Table 4.5.

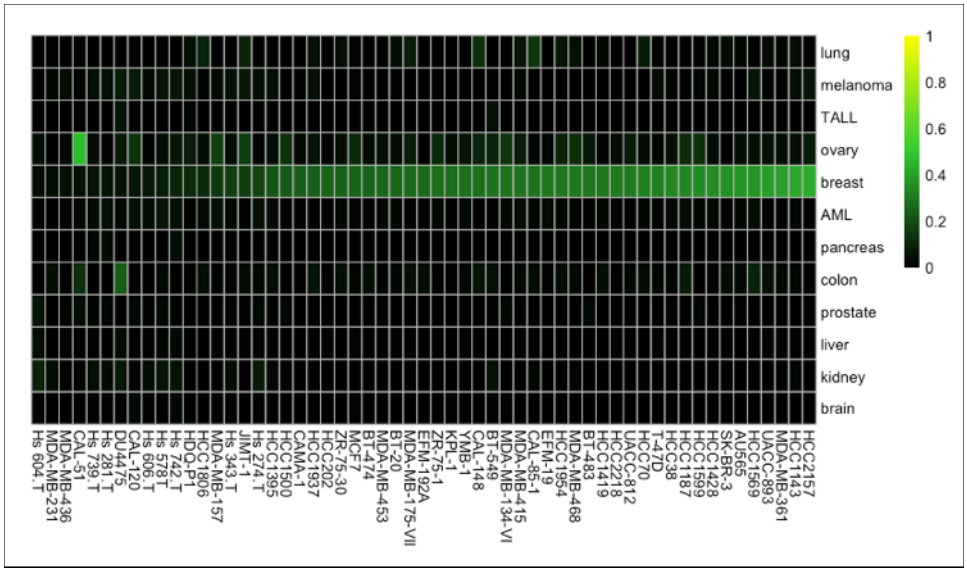


Figure 4.10 – Classification of breast cancer cell lines

Analyzing the expression data from the 57 breast cell lines through Cancer CellNet resulted in low- to mid-range probability scores for classification into breast, as shown in Figure 4.10. The cell line depicting maximum likelihood to the breast cancer classifier is HCC2157 with a score of 0.43, and this cell line has very few citations on PubMed. A few of these cell lines show almost zero probability for classifying into a breast derived cell line and this could be due to the fact that most of these cell lines were not derived from the primary tumor tissue, but further analysis is required before the causality for poor classification can be determined. MCF-7 has been used extensively in clinical research and is the most cited cancer cell line across all tumor types, even though its classification score was only ~0.25. This line

was especially useful for the development of breast cancer drugs like Tamoxifen, as it proves to be a reasonable drug target analog for primary breast tissue.

It can be seen from the classification heatmap that many breast cancer cell lines show an overlap with ovarian classification to some degree. In some cases, the ovarian classification scores exceed that of breast, although as the classification into breast increases, the ovarian non-specificity goes down. In order to understand the extent of breast-tumor GRN establishment in these cell lines into breast, their GRN status was measured with respect to both ovary and breast and the results are depicted in Figure 4.11.

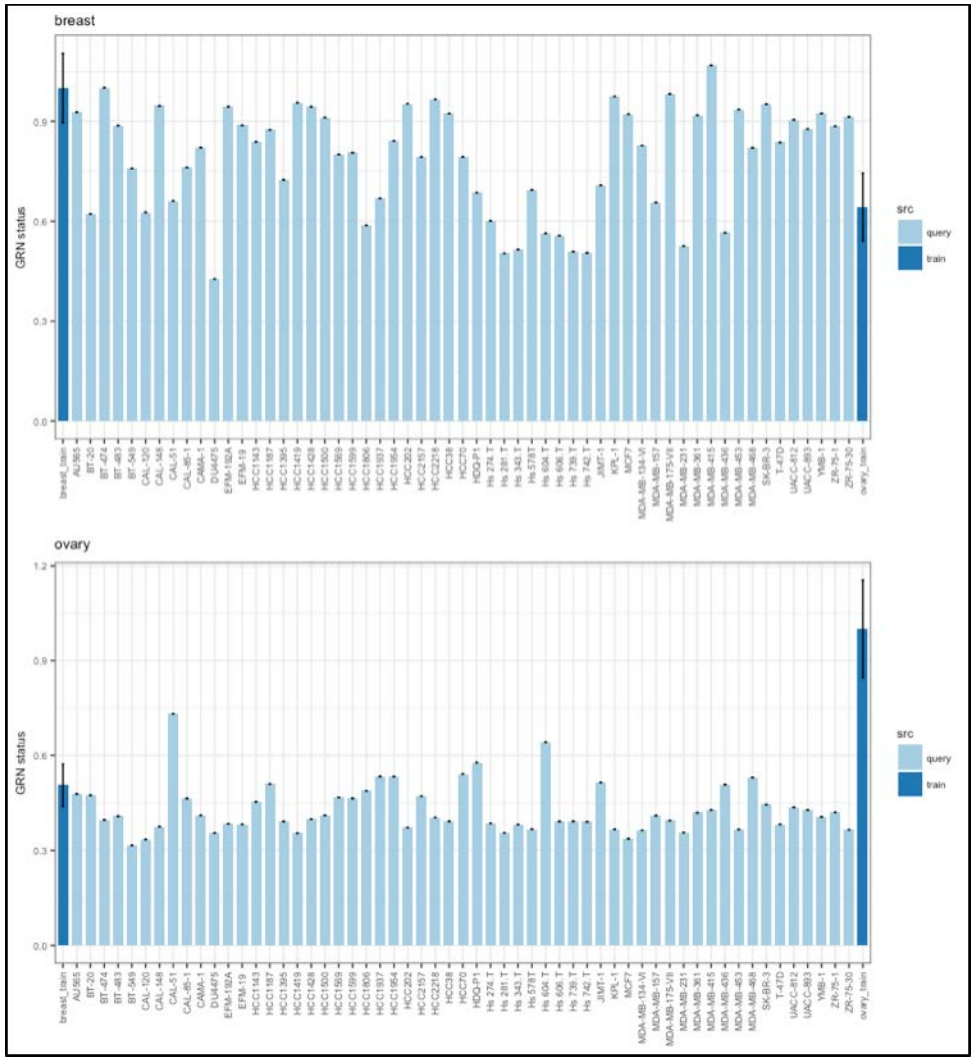


Figure 4.11 – GRN status analysis for breast cancer cell lines

The Breast panel in Figure 4.11 depicts the breast-tumor GRN status for all the breast cancer cell lines, and on the extremes are the GRN status for the primary tumor data sets. The ovary training dataset shows more than 60% similarity to the breast GRN even though only unique interactions are classified as a part of a ttGRN, during CLR based GRN construction during CellNet training. This implies that the inherent GRNs for these two tissues share a high degree of similarity even though identical players are not involved in their construction. However, while probing the ovarian GRN status (Ovary panel) for breast cancer cell lines, only a few cell lines like CAL-51 and Hs604.T rise above the breast primary tumor training threshold and show some degree of ovarian classification. These are the same cell lines that also show poor classification in the heatmap represented in Figure 4.10.

4.2.7 Prostate cancer cell lines

The prostate gland is present as a part of the male reproductive system, and is located around the bladder, wrapped around the urethra as it exits the bladder. Cancer in the prostate is the most widely occurring cancer in men and is typically diagnosed in men over the age of fifty. This is one of the most extensively researched cancers and there are several known genetic and environmental causes that may trigger the onset of prostate cancer. The most common genetic factors are RNASEL, BRCA1, BRCA2 and HoxB13[76].

The most widely used and cited cell line models for prostate cancer are DU-145, LNCaP and PC-3, and were considered the gold standard for cell lines of prostate cancer[77]. However, as can be seen from our classification metric depicted in Figure 4.12, they are not the best classified primary prostate tumor-like cell lines. DU-145 was derived in 1971[78], from a brain metastatic prostate tumor that showed hormone independence as the cells do not express Androgen Receptors (AR). PC-3 was isolated from a vertebral metastasis of the prostate tumor in the late 1970s[79] and is similar to DU-145 with respect to hormone

intolerance. The third cell line, LNCaP was also isolated from a secondary metastatic site, this time the lymph node[80], but this cell line shows AR expression and hence, is responsive to hormonal treatments. Since primary prostate tumors show expression of ARs, it is interesting to note that LNCaP resembles primary tumors better than the other two gold standard cell lines.

The three most cited cell lines used in prostate cancer, were isolated from metastatic tumors and hence, are by nature, more aggressive and lose some of the characteristics of the parent tumor. Even though they may serve as good models to test novel prostate cancer drugs, it becomes imperative to have a good cell line model to represent the primary tumor, and currently, all cell lines derived from prostate cancer, are from metastatic sites and there is a need to isolate cells from the primary tumor site, immortalize them and maintain them in culture. Studies with these cell lines will help prevent remission in the parent site as well, in cases where the prostate cancer is not completely surgically excised.

The two cell lines that show maximum similarity to primary prostate tumors are VCaP and MDA PCa 2b. Both these cell lines were isolated from vertebral metastatic sites but they have unique distinguishing features that places them closest to the primary prostate cancer cells. VCaP cells exhibit a unique gene rearrangement in serine 2–ETS regulated gene (TMPRSS2-ERG)[81], which is a transmembrane protease that plays an important role in early prostate cancer invasion and metastasis. This rearrangement results in the creation of an androgen responsive oncoprotein and this cell line is one of the very few cell lines that contains this feature. MDA PCa 2b is also unique in the fact that it was isolated from an African American patient and this becomes important because of the fact that prostate cancer is most common among men of African American descent[82], and this incidence dominance may have affected the skewing of the primary tumor dataset as well. In cases like this, where there is a

known predisposition for members of a certain section of the population for certain forms of cancer, drug development needs to be focused on developing targeted therapies and the choice of the right cell line becomes crucial in that process.

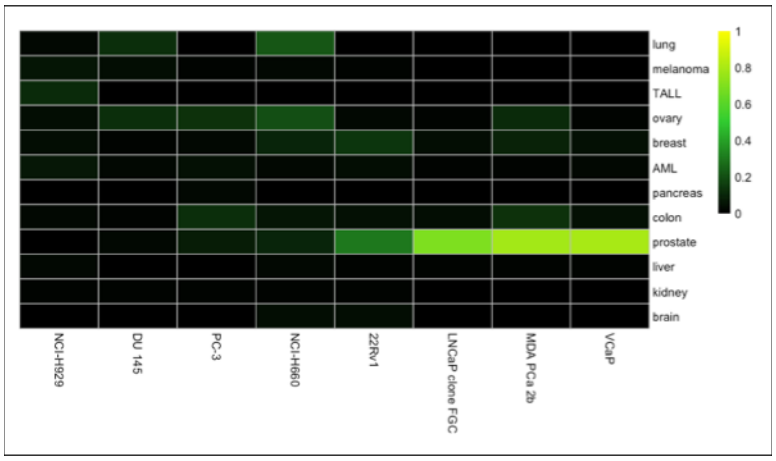


Figure 4.12 – Classification of prostate cancer cell lines

Among the 8 cell lines profiled in this study, 5 of them show minor classification into breast or ovary or both, and as prostate cancer is a disease that occurs in males, the overlap with female organs like breast and ovary was surprising. However, delving into the genetic mutations that are present in prostate revealed that the BRCA1/2 genes[83], which are causative mutations in both breast and ovarian cancers, is also predominantly mutated in prostate cancer. Thus, the non-specific classification into breast/ovary that we observe in our query may be a result of these gene networks, that play a role in all three cancers.

4.2.8 Conclusions from this section

The classification heat maps for individual tumor types described above, provide an understanding into the relative levels of likelihood for a tumor derived cell line to resemble the in-vivo tumor with respect to its GRN profile. Overall, the number of cell lines derived from a particular tumor type that resemble the parent tumor are a relatively small percentage of the total number of cell lines derived from the same tumor. This is depicted in Figure 4.13,

where the cell lines derived from the tumor types described above are plotted along the x-axis on the basis of their classification score. Almost 70% of cell lines plotted show a 50% or lower classification score, with respect to their corresponding in-vivo tumor counterparts. Some of these cell lines with low scores may have specific properties that make them attractive models for certain studies, for ex- MCF-7, although is a low scoring breast line, expresses the hormone receptors and thus is a useful model in drug screening studies. However, the characteristics of the cell line used to study the properties of the tumor biology itself or those used as precursors to clinical trials need to resemble their in-vivo counterparts to a reasonable extent and obtaining their probability score using GRN profile can serve as one such screening tool to prune cell lines that are unfit models for cancer. This can also serve as an important parameter to record while isolating/establishing a new cell line from novel or rare cancer types.

Also shown in Figure 4.13 is the relationship between the classification score obtained from Cancer CellNet with Normalized citation index. This term is defined as follows:

$$\text{Normalized citations index} = \frac{\text{Number of citations for cell line}}{\text{Number of citations for tumor type}}$$

The ideal relationship between these two parameters would be a direct correlation but the figure shows an almost inverse correlation, with a few exceptions. This implies that in majority of cases, the most heavily cited and hence most researched cell line for a particular tumor type has low resemblance with that primary tumor. 13 out of 19 cell lines that have a normalized citation index above 0.1 fall below the 50% classification score threshold. Similar results were observed by Domcke et al when they did this analysis for ovarian cell lines, where popular cell line models do not recapitulate tumor characteristics. These results together reiterate the need for a thorough re-characterization of cell line models for primary tumors prior to use for

academic or medical research, using high throughput genomics and proteomics techniques available now and this platform can be one such tool to aid in this process.

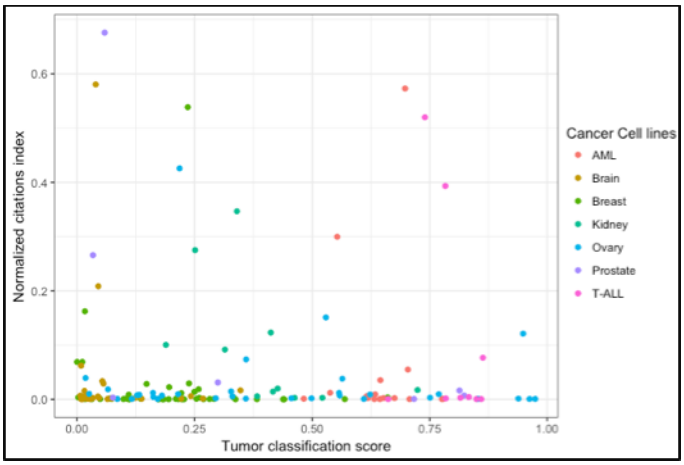


Figure 4.13 – Correlation between tumor classification scores and normalized publication index

4.3 Querying cell lines from tumors not in the training dataset

Tumors exhibit heterogeneity in a cellular, physiological and genetic level[84] and it is this property that makes it a very hard disease to cure completely. Metastatic cancers result in formation of tumors in secondary sites and these masses exhibit properties that are sometimes an amalgamation of both primary and secondary tissue sites.

The CCLE database contains cell lines derived from tumor types that are not a part of our training data set. In order to test the similarity of cell lines from different tumors to the primary tumors that are part of our training data, the ttGRNs derived from these cell lines were queried and the results for three such data sets are described below.

4.3.1 Cancer cell lines from hematopoietic lineages

Blood cancers can be of three major types: Leukemia, Lymphoma and Myeloma[85]. Leukemia is the cancer of the bone marrow and results in the production of excess of abnormal white blood cells into the bloodstream[86]. Lymphoma, as the name suggests, is the cancer of the lymphatic system, which is responsible for removal of excess fluid from the

body. Cancer in this system results in the production of abnormal lymphocytes, which are also important in fighting off infection[87]. Both these diseases decrease the immunity in the patient. Myeloma is the cancer of the plasma cells, which are white cells that produce antibodies to fight disease, and thus antibody production is impaired in patients suffering from myeloma[88].

In the CCLE database, there were 109 cell lines that were derived from hematopoietic or lymphoid tissue but were not annotated as TALL or AML, which were the two types of blood cancers that were a part of the training data set. Thus, these 109 cell lines were grouped under the broad umbrella of blood-derived cancer cell lines and queried against our Cancer CellNet derived classifiers. The resulting heat map, depicted in Figure 4.14, showed some specific classifications into TALL and/or AML but a majority of the cell lines showed no specific alignment with any one tissue type, with a reasonably high probability.

37% of them aligned with TALL, 18% with AML, 0.01% with breast and ovary each and the remaining ~45% did not show any specific tissue classification, (using a 0.2 classification threshold) as shown in Figure 4.15.

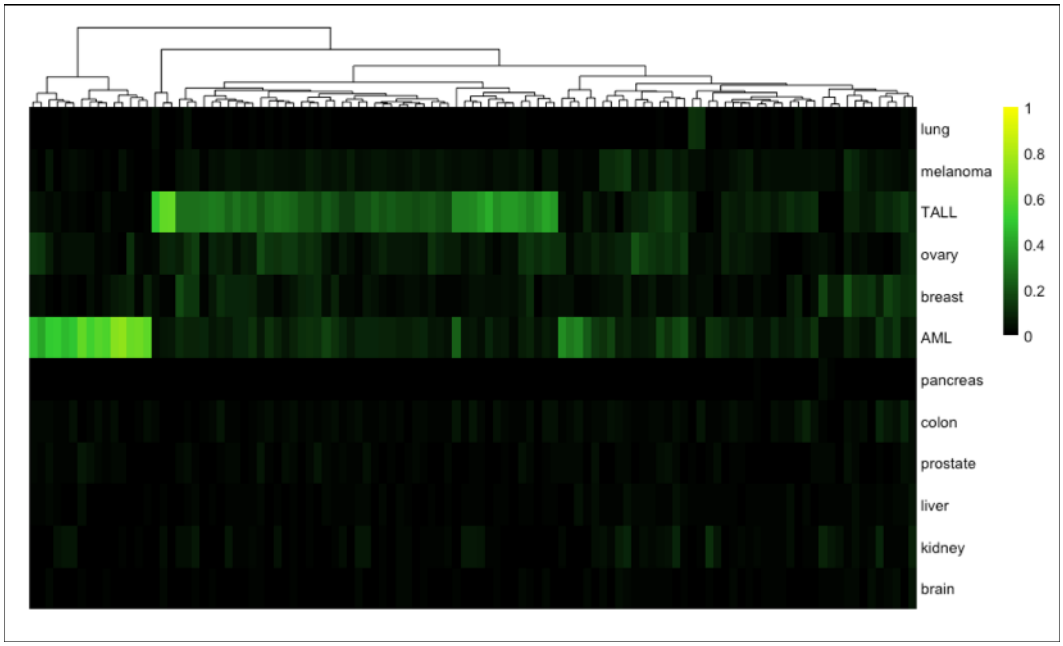


Figure 4.14 – Classification of hematopoietic cancer cell lines

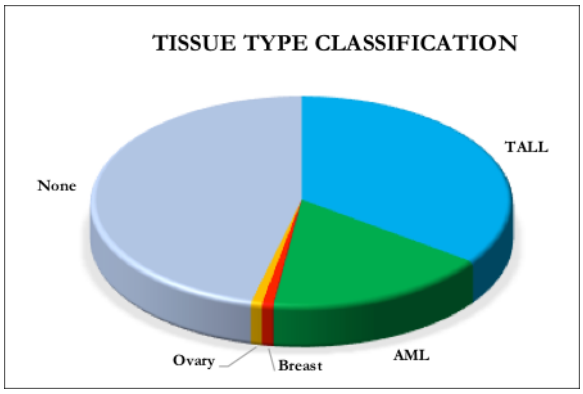


Figure 4.15 – Tissue specificity of blood-derived cancer cell lines

Analysis of the cell lines that actually aligned with either TALL or AML led to some interesting correlations. Burkitt Lymphoma is a form of non-Hodgkin's lymphoma that originates in the B cells and it is the fastest growing human tumor that leads to high fatality[89]. In our query data set, there were 10 cell lines derived from Burkitt Lymphoma and all of them classified as TALL when run through Cancer CellNet, with a probability of 28%. Among other types of cancers that also classified as TALL are Hodgkin lymphoma, chronic lymphocytic leukemia and B cell lymphoma. Cell lines that classified as AML were predominantly the cell lines derived from Chronic Myeloid Leukemia or CML, and the two diseases share more commonalities in gene expression profiles than either of them with T-ALL, although they result in different manifestations of the disease and clinical symptoms.

4.3.2 Endometrial cancer derived cell lines

Endometrial cancer arises in the endometrial lining of the uterus and is common in post-menopausal women above the age of 50. If diagnosed early, this disease has a good prognosis rate and is treated through surgical removal of the uterus (abdominal hysterectomy) but still remains the third leading cause of death in cancers that affect women[90]. Some of the main risk factors for the occurrence of this cancer are obesity, estrogen exposure and genetic predisposition. The estrogen dependence becomes an important point of consideration as

selective estrogen receptor modulators (SERMs), like Tamoxifen, are commonly used treatments for breast cancer, and thus making the incidence of endometrial cancer higher among patients who have been treated with SERMs for breast cancer[91]. In many cases, endometriosis is also found to be a precursor for ovarian cancer in women.

There are two main types of endometrial cancer, Type-I which occurs in 80% of cases and is endometroid and hyper-estrogenic in origin. Type-II on the other hand, is not endometroid in origin and does not show association with hyper-estrogenic factors like obesity and is usually highly dangerous with great metastatic and recurrence potential[91].

There are 25 cell lines that are of endometrial-tumor origin in the CCLE database, which have been included in this analysis. Figure 4.16 shows the heatmap for the classification scores for all cell lines, arranged in increasing order of classification score with ovary ttGRN.

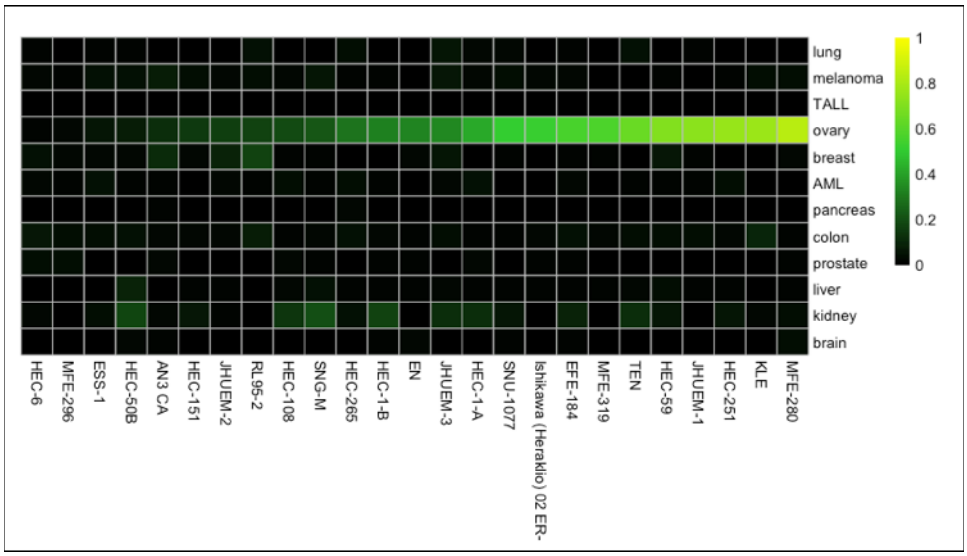


Figure 4.16 – Tissue specificity of endometrial cancer cell lines

As can be seen, the cell line showing highest similarity and least non-specificity is MFE-280, which had been isolated in 1997[92] but only been cited in literature less than 10 times. More than 60% of these cell lines have a 25% or above probability of classifying as an ovarian cancer, which is not surprising considering the similarity in tissue composition and

molecular players between the ovary and the endometrium. The ovarian tumor dataset comprised of ~20% of tumors that were of endometroid origin[93] and this could also have led to the similarity of classification seen by endometrial cancer cell lines to ovarian tumors.

Some cell lines show a minor classification with kidney but the probability score is less than 0.2 and is not considerably significant. Zhou et al, in 2007, had published a comprehensive study researching the properties of 16 endometrial cancer cell lines[94], some of which overlap with our study, but they do not compare the cell lines to the primary tumor to ascertain molecular and phenotypic similarities. The study however, catalogs properties of these cell lines such as proliferation rate, morphology, hormone receptor expression, tumorigenicity etc. in detail.

The cell line showing faint breast classification, RL95-2, was analyzed for the expression of Estrogen Receptor (ER) but all endometrial cancer cell lines showed a uniform expression of ER. The expression of the ten breast cancer associated TFs[95], namely, SOX10, NFATC2, ZNF354C, ARID3A, BRCA1, FOXO3, GATA3, ZEB1, HOXA5 and EGR1, was looked into next and all of them, except for GATA3 were found to be near basal levels. GATA3, however, showed a three-fold and this could have resulted in breast-like classification. MFE-296, which is one of the cell lines that shows negligible classification as primary endometrial cancer even though it was originally obtained from differentiated human endometrial adenocarcinoma. This is one of the few endometrial cancer cell lines to show tetraploidy[96]. In-vivo analysis of tetraploid endometrial tumors reveal that they are similar to diploid tumors, but the poor classification of this cell line with our primary tumor data set shows that further analysis needs to be conducted into understanding the mechanisms in which ploidy affects cancer characteristics in the endometrium.

4.3.3 Bone cancer derived cell lines

Primary bone cancers are a subtype of a group of cancers called sarcomas, which broadly refers to the cancers that start in the bone, muscle, connective tissue, fat or blood vessels[97]. Bone cancers are of many different types and the cell lines we have in our query dataset belong to 4 distinct types – Ewings sarcoma, osteosarcoma, chondrosarcoma and giant cell tumors of the bone. The most common among these, osteosarcoma, is characterized by the production of calcified bone (osteoid) by malignant cells, in an aberrant fashion, leading to carcinogenesis. Chrondrosarcoma is a cancer of the cartilage tissue while Ewings sarcoma occurs in soft tissue. Since osteosarcoma affects children and young adults, a lot of attention has been paid to this tumor type, by researchers worldwide[98].

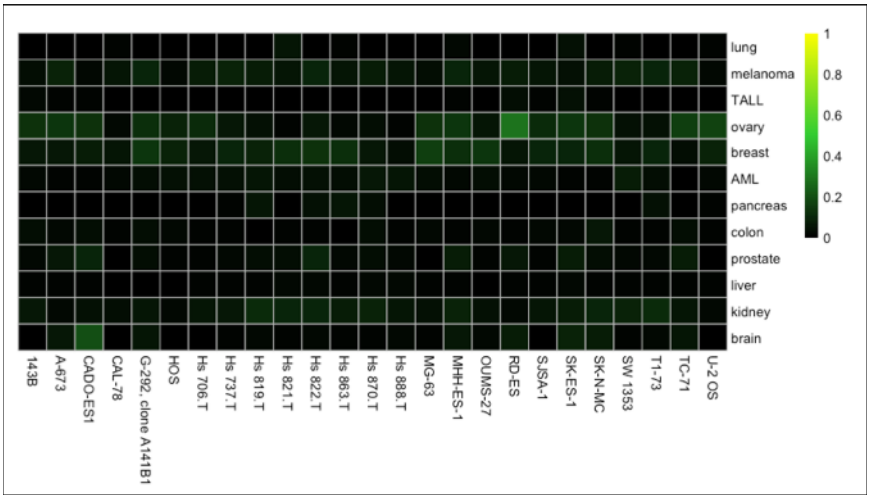


Figure 4.17 – Tissue specificity of bone-derived cancer cell lines

Analyzing the bone cancer derived cell lines as a part of this study resulted in a very non-specific heat up as shown in Figure 4.17. The tissue composition and the genetic networks involved in bone cancer formation are starkly different as compared to other tissues, that are a part of the training data. In the heat map, there is low background for almost all cell lines that shows poor classification. Without a bone cancer training dataset, it becomes difficult to analyze the fidelity of osteosarcoma cell lines, and this issue is addressed in section 4.4.

4.3.4 Conclusions from this section

The intended advantage of a tool like Cancer CellNet is the ability to classify an unknown sample into a known tumor type. When querying Cancer CellNet with cell lines derived from tumors that are not a part of the training dataset, one of two things can happen. If the query data resembles one or more of the training tumor types to a reasonable degree, like the similarity between endometrium and ovary, they classify as that tumor type. However, while querying a completely unrelated tumor type like bone, with these classifiers, the resulting heatmap shows poor classification scores with all tumors and becomes hard to interpret, thus requiring CellNet to be retrained including necessary tumor type controls. This calls for the need to ensure maximum diversity within the training dataset, so characterization of unknown samples becomes possible.

4.4 Re-training CellNet with osteosarcoma tumor samples

In order to be able to analyze the osteosarcoma derived cell lines mentioned above, CellNet was retrained with all of the primary tumor samples listed in Table 3.1, and 48 osteosarcoma primary tumor expression datasets. This allowed the construction of bone tumor type ttGRN based classifiers and re-construction of the same for the other tissue types. We also made the GRNs larger, allowing more candidate TFs to be a part of each ttGRN in order to construct a more comprehensive classifier, which will be used for further analysis.

As described in section 4.1, prior to using ttGRN based classifiers generated by CellNet, it becomes crucial to ensure that the classifiers are specific and are capable of classifying query data satisfactorily. This is done by the "split and assess" method, described above. The results of this analysis are shown in Figures 4.18 and 4.19 below.

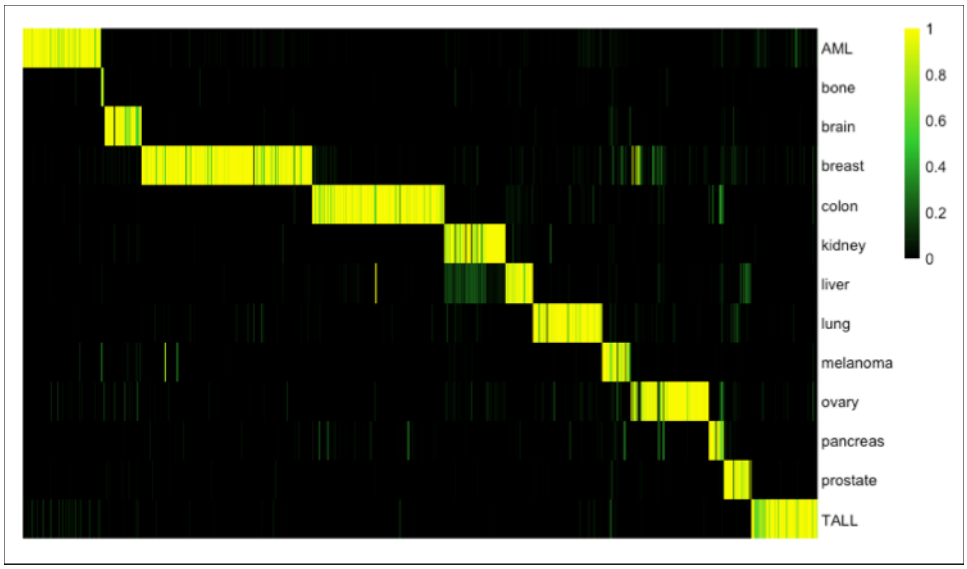


Figure 4.18 – Classification scores to test the quality of the new training classifiers

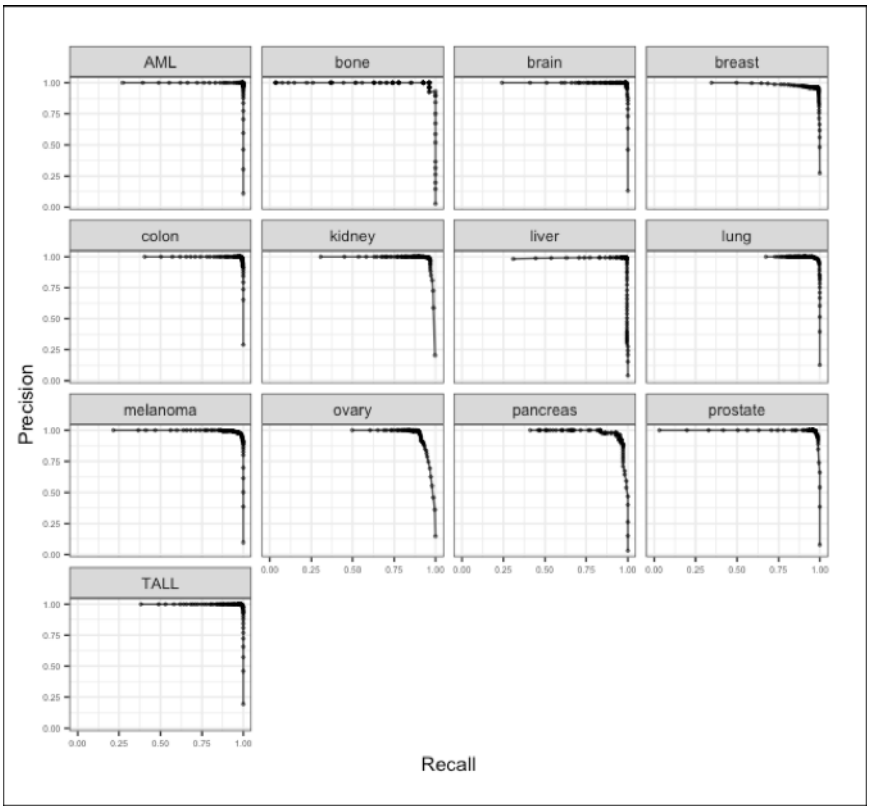


Figure 4.19 – Precision-Recall curves for the new CellNet generated classifiers

Although the classifiers are successful in classifying the tumor data in accordance with their tumor type of origin, there appears to be some background classification for tumor types

like kidney, pancreas, melanoma and ovary and most of this non-specific background is appearing to be with breast or liver tumor type classifier. The PR curves are assessed to better understand the sensitivity and precision of these classifiers to accurately predict the tumor type of origin for query data.

Similar to Figure 4.2, ovary and prostate show a less-than-perfect PR curve, but the precision is high with a sensitivity of about 0.9, thus making the classifiers a good standard. The bone classifier performs a little poorer in comparison, but that may be due to the relatively smaller number of samples that constitute the training data set. With these classifiers, the osteosarcoma cell lines were re-analyzed and the results are shown in Figure 4.20, where the cell lines are arranged in increasing probability to classify as bone tumor type.

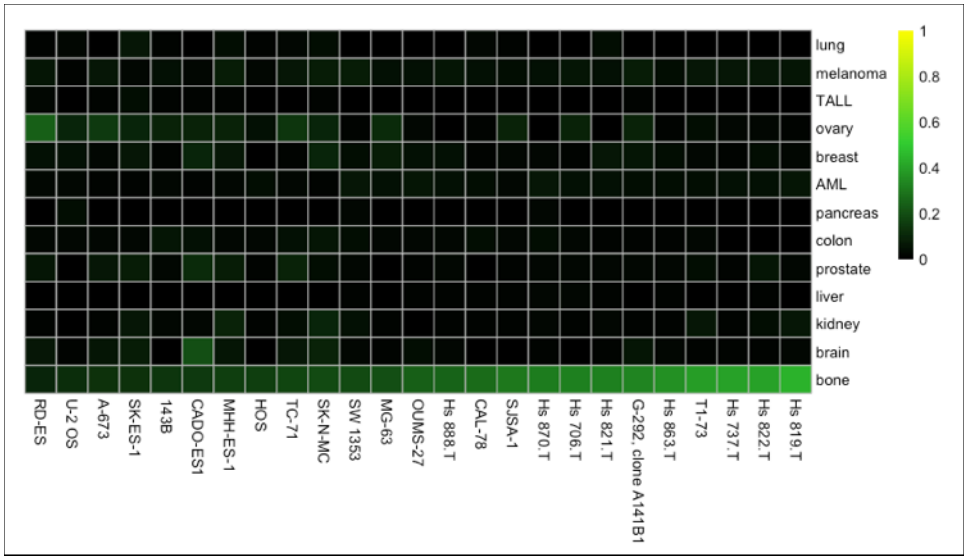


Figure 4.20 – Tissue specificity of bone-derived cancer cell lines using new classifiers

A study by Mohseny et al in 2011 characterized the in-vitro differentiation and in vivo tumorigenic potential for 19 osteosarcoma cell lines, out of which only 5 overlap with our study[99]. They identify 8 cell lines which have the ability to form tumors in immuno-compromised mice, one of which (HOS143B) was identified as a highly metastatic tumor-forming cell line. However, the tumors formed were not analyzed with respect to traditional

osteosarcoma expression profiles, to ensure that the tumor formed recapitulates the features of the cancer that the cell lines were derived from. The cell lines that were ranked by this study to be good representatives for OS, based on these two experiments, were among those that classified as poor models in our study. U2OS, which is one of the oldest human cell lines established is also having low/poor classification to primary tumor data. These differences between quantitative and qualitative analysis of fidelity of cell lines show the need for more stringent characterization techniques, and we believe that Cancer CellNet is a step in that direction.

4.5 Identification of the unknown origins of certain cancers

"Cell of origin" are those cells that receive the first genetic hit or hits, that contributes to the initiation and in some cases, the propagation of the cancer[100]. There are certain cancers where these cells of origins are not identified and using Cancer CellNet, we aim to identify the possible cells from which these cancers could arise from.

4.5.1 Tissue origins of cancers with unknown primary tumors

The region or tissue of the body where cancer originates is called the primary site and that tumor is called the primary tumor. Sometimes, this tumor metastasizes into other organs and these are called secondary tumors[101]. Secondary tumors exhibit characteristics of the tissue where the primary cancer originated and are hence, named after that tumor. However, around 2 - 5% of all patients have metastatic tumors where the primary site for the tumor remains unknown and these cancers are called Cancers with Unknown Primaries (CUP)[102]. This usually occurs if either the primary tumor is too small to be identified and/or if the primary tumor regresses after metastasis and hence, is unidentifiable. Although the metastasis for CUP can be found in any part of the body, the most common tissues are lymph nodes,

liver, lungs and skin[103]. There are four major types of CUPs based on the cells present in the metastatic tumor – adenocarcinoma, squamous cell carcinoma, neuroendocrine carcinoma and poorly differentiated tumors, among which adenocarcinomas comprise of 60% of all diagnosed CUPs[104].

With advances in molecular imaging and genomics, there have been significant progress made in identifying the mechanisms by which the secondary tumor keeps the characteristics of the primary tumor obscure. However, with Cancer CellNet, the GRN status of these secondary tumors can be quantified and compared across our training tumor types to obtain similarity measures. The study by Kurahashi et al in 2013 generated microarray expression profiles for 60 such CUP cancers obtained from patients[105] and we use this data set as a query into Cancer CellNet. The results are depicted in Figure 4.21.

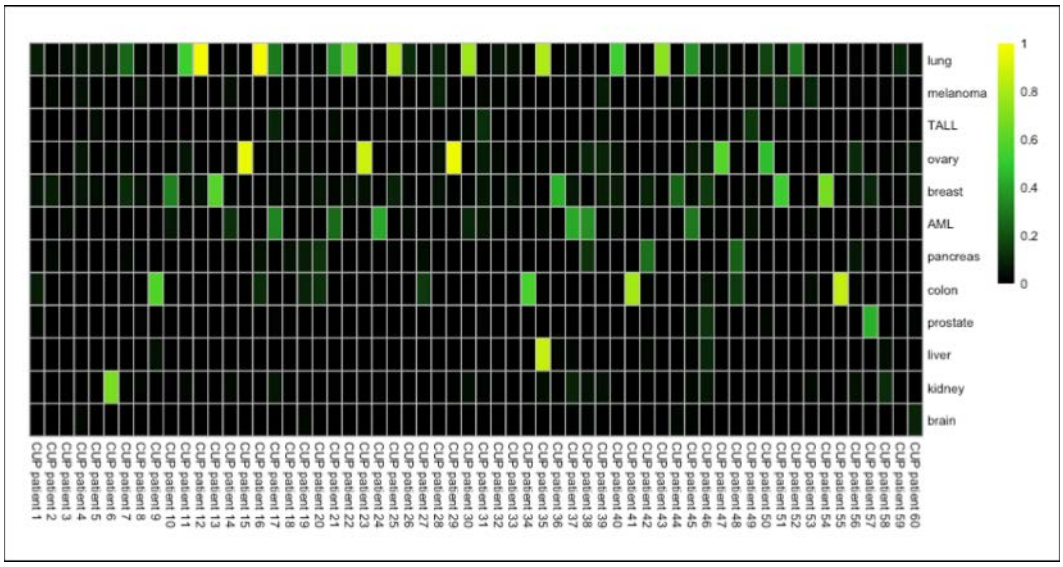


Figure 4.21 – Classification score heatmap – Cancers of Unknown Primary samples

This dataset was collated by isolating CUPs from the lymph nodes or ascites fluid of 60 patients and their expression profile was obtained. As can be seen from the figure, around 20% of the samples show maximum similarity to classify as a lung tumor, while 10% classify

as ovarian. Interestingly, CUP from patient 35 showed higher than 70% classification with both lung and liver, which do not share a lot of functional, phenotypic or GRN similarity.

A representation of the tumor types the CUP samples classified as and their relative distribution, is shown in Figure 4.22. There seem to be some tissues, like melanoma, prostate and brain, which have no specific and very minimal non-specific classification. This information will be useful for clinicians to direct them to look at tissues like lung, which have a higher probability of having had the primary tumor, which may/may not be visible.

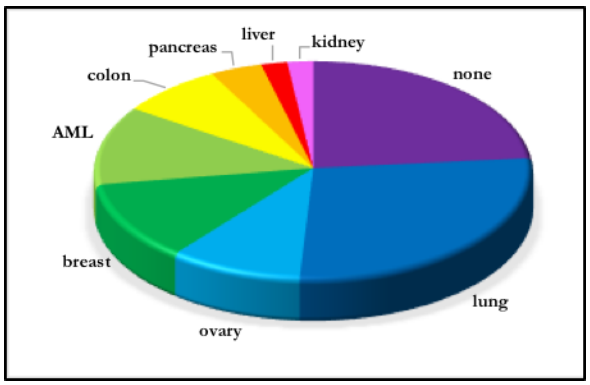


Figure 4.22 – Putative tissues of origin of CUP samples

Although Cancer CellNet classifiers are able to discern the putative origin for a majority of the samples profiled based GRN similarity with the ttGRNs, around ~15% of these samples still remain non-specific with respect of primary tumor origin. However, inclusion of additional primary tumor datasets to train CellNet, like thyroid, gastrointestinal system etc. may help provide additional tumor profiles to compare against.

We propose the possible use of this platform as a diagnostic tool for cases where the primary cancer sites are unidentifiable or or is unclear, as it provides a quick and quantitative understanding of the tumor profile and can direct the treatment methodologies in those cases.

4.5.2 Analyzing the origins of Merkel Cell Carcinomas

Merkel Cell Carcinoma (MCC) is a highly aggressive form of skin cancer and in approximately 80% of the cases is caused by Merkel cell polyomavirus (MCV). It is also called neuroendocrine carcinoma of the skin and predominantly occurs among older people, due to prolonged sun exposure or a weakened immune system[106]. A recently discovered virus called Merkel Cell Polyomavirus (MCV) is shown to contribute to the development of MCC[107] but the cells where the disease originates is still unknown and is an area of active research.

Long term resident cells or epidermal stem cells, as they are popularly called, are present in the inter-follicular epidermis and are the cells of origin for MCC in mice[108]. The figure below, adapted from Tilling et al shows the potential cells of origin for MCC in humans, and their putative mechanisms for carcinogenesis into MCC[16].

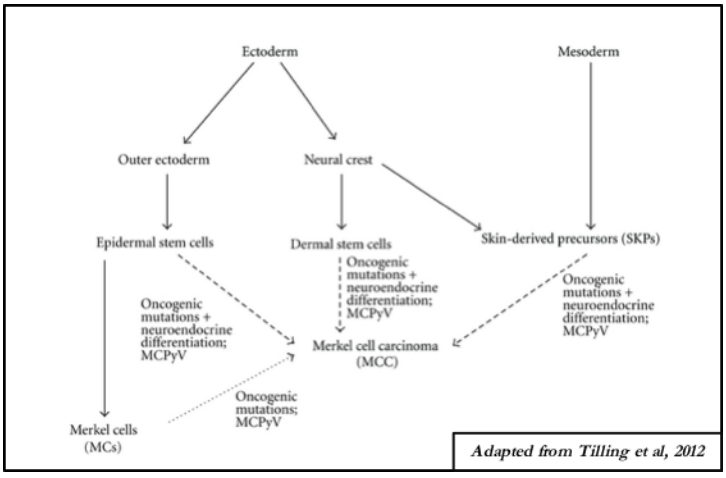


Figure 4.23 – Putative cells of origin and paths to the formation of MCC

In order to understand the tumor type similarity between MCC and the primary tumor types in our training data set, the data obtained by Harms et al (GEO - GSE39612) was used as the query dataset to run through Cancer CellNet[109]. The results are shown in Figure 4.24. Since the primary dataset does not include a neuroendocrine tumor profile, most of the MCC samples align with brain with scores ranging from 20% to 45%. Tumor profiles between

metastatic tumors and primary tumors do not show any significant differences with respect to classification scores. Dermal stem cells, which are considered one of the cell types that is thought to lead to the formation of MCC, are known players in the formation of melanoma. A very small subset of the patient samples shows a classification as melanoma, and all except one among those, have a significant probability score. Interestingly, this sample did not classify as brain, exhibiting one of the lowest score among these samples.

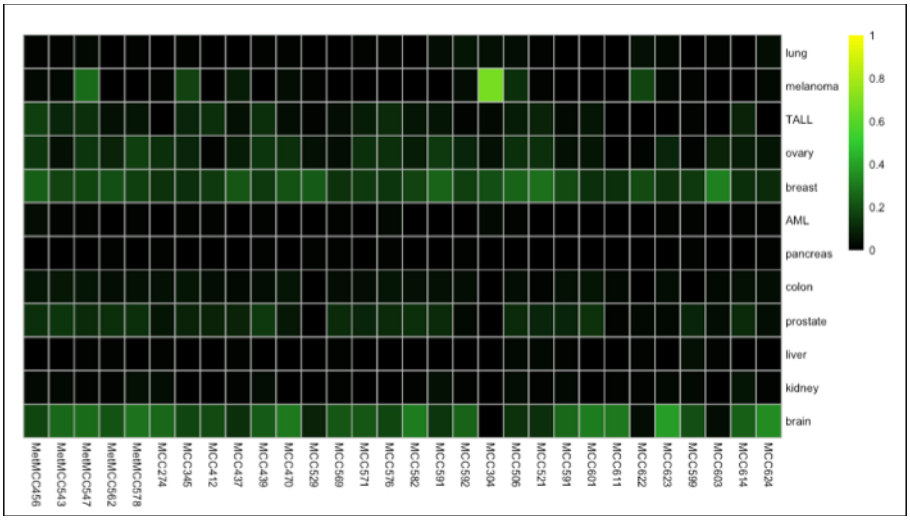


Figure 4.24 – Classification score heatmap – Merkel Cell Carcinoma samples

MCV is a known contributor for the propagation of the MCC and the MCV status of these samples was probed into. 13 of the samples had tested positive for the viral status, as determined by PCR for the viral gene. Although the number of metastatic tumors included in this study are only 4, all of them tested negative for the viral genes and this correlation is yet to be explored. MCV is also known to infect stem cells, and Merkel cells, which were originally considered a frontrunner as the cell of origin, have been shown to not exhibit the presence of this virus due to its post-mitotic nature.

The results obtained from Cancer CellNet may be the quantitative tool the field lacked to analyze the tumor profile of MCCs and with the appropriate tumor classifiers and larger number of patient samples, the tumor type similarity for MCC can be identified.

Chapter 5 - Discussion

5.1 Conclusions

Using Cancer CellNet as a tool, the cell lines from ten different tumor types were analyzed and ideal in vitro models were proposed based on the classification scores. While querying cell lines derived from tumors that are not a part of the training data, CellNet is able to predict the closest tumor type as the tissue or origin. When none of the classifiers resembles the query data, poor classification scores are encountered against all tumor types and in these cases, Cancer CellNet needs to be retrained with appropriate primary tumor expression data.

The results obtained from this project demonstrate the use of Cancer CellNet platform as a diagnostic tool in the clinical setting. While querying clinical samples obtained from patients, where the cell origins of the cancer is unknown, based on the classification score with the ttGRNs, we can identify the putative tumors of origin to a reasonable degree. As cancer is a heterogeneous disease, oftentimes there is a chance for a misdiagnosis while using qualitative assays but since Cancer CellNet is a quantitative tool and the results shown above prove minimal overlap with other tissues, accuracy of diagnosis can be significantly improved. Since microarrays are quite inexpensive and obtaining expression data from samples is faster than a majority of clinical assays, using Cancer CellNet in a diagnostic setting to accurately predict the cancer profile has tremendous market potential.

5.2 Limitations of the study

Although the advantages of Cancer CellNet are many, we acknowledge that our study has a number of limitations. These are listed below:

1. The extent of diversity in the training data is limited and in order to be able to successfully handle samples from any kind of cancer, other tumor types like GI,

cervical, thyroid, and individual types of blood cancers like lymphomas, myelomas etc. need to be added. Improving the specificity of the classifiers will vastly increase the diagnostic value of CellNet in screening samples and identifying tumor type.

2. In this study, one dataset per cell line was used to determine the fidelity. However, in order to predict efficient in-vitro models for cancer with high confidence, more replicates per cell line need to be incorporated and the analysis repeated.
3. Cancer being a heterogeneous disease, has multiple cells that are involved in the diseased tissue and sometimes, also contribute to the disease initiation and progression. Performing microarray on clinical samples may be an effective way to identify an overall cancer profile but in order to use Cancer CellNet as a tool for investigating the cell types of origin, it becomes more efficient to use expression data from single cells as opposed to a collection of cells. In this respect, modifying Cancer CellNet to be able to analyze single cell RNA-sequencing data, which provides the maximum resolution into the transcriptional state of a cell presently, will make more leeway into understanding the cell origins of cancer.
4. The number of tumor samples used per tumor type varies, but since CellNet uses a modified CLR algorithm in its GRN construction, a minimum of 60 samples per tumor type is required to construct a classifier with accurate ttGRN profiles. When it comes to using CCN for rare tumor types, obtaining this data set might be challenging.

5.3 Future directions

Adapting Cancer CellNet with additional primary tumor types and querying them against multiple datasets obtained from the model cell lines predicted by this study, will enable us to predict cell line models for primary tumors with higher accuracy and we intend to do this with microarray and RNA-seq data to obtain a better resolution into cancer heterogeneity.

Chapter 6 – References

1. American Cancer Society. Cancer Facts and Figures 2016 Atlanta: American Cancer Society. 1–72 (2016).
2. American Association for Cancer Research. AACR Cancer Progress Report 2016. *Clin Cancer Res ;* 1–143 (2016).
3. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2016. *CA: A Cancer Journal for Clinicians* **66,** 7–30 (2016).
4. Gillet, J.-P., Varma, S. & Gottesman, M. M. The Clinical Relevance of Cancer Cell Lines. *J Natl Cancer Inst* **105,** 452–458 (2013).
5. Borrell, B. How accurate are cancer cell lines? *Nature* **463,** 858 (2010).
6. Holliday, D. L. & Speirs, V. Choosing the right cell line for breast cancer research. *Breast Cancer Res.* **13,** 215 (2011).
7. Masters, J. R. Human cancer cell lines: fact and fantasy. *Nature Reviews Molecular Cell Biology* **1,** 233–236 (2000).
8. Nelson-Rees, W. A. The identification and monitoring of cell line specificity. *Prog Clin Biol Res* **26,** 25–79 (1978).
9. McLendon, R. *et al.* Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455,** 1061–1068 (2008).
10. Bignell, G. R. *et al.* Signatures of mutation and selection in the cancer genome. *Nature* **463,** 893–898 (2010).
11. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463,** 899–905 (2010).
12. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483,** 603–607 (2012).
13. Domcke, S., Sinha, R., Levine, D. A., Sander, C. & Schultz, N. Evaluating cell lines as tumour models by comparison of genomic profiles. *Nature Communications* **4,** 55 (2013).
14. Bulyk, M. L. & Walhout, A. J. M. in *Handbook of Systems Biology* 65–88 (Elsevier, 2013). doi:10.1016/B978-0-12-385944-0.00004-6
15. Pavlidis, N. & Fizazi, K. Carcinoma of unknown primary (CUP). *Critical Reviews in Oncology/Hematology* **69,** 271–278 (2009).
16. Tilling, T. & Moll, I. Which Are the Cells of Origin in Merkel Cell Carcinoma? *Journal of Skin Cancer* **2012,** 1–6 (2012).
17. Vargo-Gogola, T. & Rosen, J. M. Modelling breast cancer: one size does not fit all. *Nature Reviews Cancer* **7,** 659–672 (2007).
18. van Staveren, W. C. G. *et al.* Human cancer cell lines: Experimental models for cancer cells in situ? For cancer stem cells? *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer* **1795,** 92–103 (2009).
19. Lum, D. H., Matsen, C., Welm, A. L. & Welm, B. E. *Overview of Human Primary Tumorgraft Models: Comparisons with Traditional Oncology Preclinical Models and the Clinical Relevance and Utility of Primary Tumorgrafts in Basic and Translational Oncology Research*. *Current Protocols in Pharmacology* **69,** 14.22.1–14.22.9 (John Wiley & Sons, Inc., 2012).
20. Richmond, A. & Su, Y. Mouse xenograft models vs GEM models for human cancer therapeutics. *Disease Models and Mechanisms* **1,** 78–82 (2008).
21. Katt, M. E., Placone, A. L., Wong, A. D., Xu, Z. S. & Searson, P. C. In Vitro Tumor

Models: Advantages, Disadvantages, Variables, and Selecting the Right Platform. *Front. Bioeng. Biotechnol.* **4,** 311 (2016).
22. DeRose, Y. S. *et al.* Tumor grafts derived from women with breast cancer authentically reflect tumor pathology, growth, metastasis and disease outcomes. *Nature Medicine* **17,** 1514–1520 (2011).
23. Scherer, W. F., Syverton, J. T. & Gey, G. O. Studies on the propagation in vitro of poliomyelitis viruses. *J. Exp. Med.* (1953).
24. Ferreira, D., Adega, F. & Chaves, R. *The importance of cancer cell lines as in vitro models in cancer methylome analysis and anticancer drugs testing.* (2013).
25. Nelson-Rees, W. A. & Flandermeyer, R. R. Inter- and intraspecies contamination of human breast tumor cell lines HBC and BrCa5 and other cell cultures. *Science* **195,** 1343–1344 (1977).
26. MacLeod, R. A. *et al.* Widespread intraspecies cross-contamination of human tumor cell lines arising at source. *Int. J. Cancer* **83,** 555–563 (1999).
27. Furlong, M. T., Hough, C. D., Sherman-Baust, C. A., Pizer, E. S. & Morin, P. J. Evidence for the Colonic Origin of Ovarian Cancer Cell Line SW626. *J Natl Cancer Inst* **91,** 1327–1328 (1999).
28. Rowehl, R. A. *et al.* Establishment of Highly Tumorigenic Human Colorectal Cancer Cell Line (CR4) with Properties of Putative Cancer Stem Cells. *PLoS ONE* **9,** e99091 (2014).
29. Gillet, J.-P. *et al.* Redefining the relevance of established cancer cell lines to the study of mechanisms of clinical anti-cancer drug resistance. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 18708–18713 (2011).
30. Langdon, S. P. Characterization and authentication of cancer cell lines: an overview. *Methods Mol. Med.* **88,** 33–42 (2004).
31. Arda, H. E. *et al.* Functional modularity of nuclear hormone receptors in a Caenorhabditis elegans metabolic gene regulatory network. *Molecular Systems Biology* **6,** 367 (2010).
32. Karlebach, G. & Shamir, R. Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology* **9,** 770–780 (2008).
33. Cahan, P. *et al.* CellNet: Network Biology Applied to Stem Cell Engineering. *Cell* **158,** 903–915 (2014).
34. Radley, A. H. *et al.* Assessment of engineered cells using CellNet and RNA-seq. *Nature Protocols* **12,** 1089–1102 (2017).
35. Faith, J. J. *et al.* Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles. *PLoS Biol* **5,** e8 (2007).
36. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* **95,** 14863–14868 (1998).
37. Butte, A. J. & Kohane, I. S. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput* 418–429 (2000).
38. Breiman, L. Random Forests. *Machine Learning* **45,** 5–32 (2001).
39. Döhner, H., Weisdorf, D. J. & Bloomfield, C. D. Acute Myeloid Leukemia. *http://dx.doi.org/10.1056/NEJMra1406184* **373,** 1136–1152 (2015).
40. Ley, T. J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456,** 66–72 (2008).

41. Cancer Genome Atlas Research Network *et al.* Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med* **368,** 2059–2074 (2013).
42. Koeffler, H. P. & Golde, D. W. Human myeloid leukemia cell lines: a review. *Blood* **56,** 344–350 (1980).
43. Rosenbauer, F. & Tenen, D. G. Transcription factors in myeloid development: balancing differentiation with transformation. *Nature Reviews Immunology* **7,** 105–117 (2007).
44. Shen, Q., Zheng, H., Chen, S. H., Yang, L. & Li, Y. The Change of Elf-1 Gene Expression Level in Hematological Malignancies. *Blood* **114,** 4704–4704 (2009).
45. MacGillavry, H. D. *et al.* Genome-wide gene expression and promoter binding analysis identifies NFIL3 as a repressor of C/EBP target genes in neuronal outgrowth. *Molecular and Cellular Neuroscience* **46,** 460–468 (2011).
46. Mitsui, S., Yamaguchi, S., Matsuo, T., Ishida, Y. & Okamura, H. Antagonistic role of E4BP4 and PAR proteins in the circadian oscillatory mechanism. *Genes Dev.* **15,** 995–1006 (2001).
47. Lin, J. *et al.* Characterization of Mesenchyme Homeobox 2 (MEOX2) transcription factor binding to RING finger protein 10. *Mol Cell Biochem* **275,** 75–84 (2005).
48. Chiaretti, S. & Foà, R. T-cell acute lymphoblastic leukemia. *Haematologica* **94,** 160–162 (2009).
49. Babusíková, O. *et al.* Phenotypic heterogeneity and aberrant markers expression in T-cell leukemia. *Neoplasma* **45,** 128–134 (1998).
50. Burger, R., Hansen-Hagge, T. E., Drexler, H. G. & Gramatzki, M. Heterogeneity of T-acute lymphoblastic leukemia (T-ALL) cell lines: Suggestion for classification by immunophenotype and T-cell receptor studies. *Leukemia Research* **23,** 19–27 (1999).
51. Pegoraro, L. *et al.* A Novel Leukemia T-Cell Line (PF-382) With Phenotypic and Functional Features of Suppressor Lymphocytes. *J Natl Cancer Inst* **75,** 285–290 (1985).
52. Weitzman, J. B., Fiette, L., Matsuo, K. & Yaniv, M. JunD Protects Cells from p53-Dependent Senescence and Apoptosis. *Molecular Cell* **6,** 1109–1119 (2000).
53. Gazon, H. *et al.* Human T-Cell Leukemia Virus Type 1 (HTLV-1) bZIP Factor Requires Cellular Transcription Factor JunD To Upregulate HTLV-1 Antisense Transcription from the 3' Long Terminal Repeat. *J. Virol.* **86,** 9070–9078 (2012).
54. Ausserlechner, M. J., Bernhard, D. & Sgonc, R. p53-induced apoptosis in the human T-ALL cell line CCRF-CEM. *Oncogene* (1997).
55. Huse, J. T. & Holland, E. C. Targeting brain cancer: advances in the molecular pathology of malignant glioma and medulloblastoma. *Nature Reviews Cancer* **10,** 319–331 (2010).
56. Omuro, A. & DeAngelis, L. M. Glioblastoma and Other Malignant Gliomas: A Clinical Review. *JAMA* **310,** 1842–1850 (2013).
57. Louis, D. N. *et al.* The 2007 WHO Classification of Tumours of the Central Nervous System. *Acta Neuropathol* **114,** 97–109 (2007).
58. Louis, D. N. *et al.* The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathol* **131,** 803–820 (2016).
59. Rini, B. I. & Atkins, M. B. Resistance to targeted therapy in renal-cell carcinoma. *The Lancet Oncology* **10,** 992–1000 (2009).
60. Hoshiyama, M. *et al.* Effect of High Glucose on Nitric Oxide Production and Endothelial Nitric Oxide Synthase Protein Expression in Human Glomerular Endothelial Cells. *NEE* **95,** e62–e68 (2003).

61. Zimmermann, K. *et al.* NOSTRIN: A protein modulating nitric oxide release and subcellular distribution of endothelial nitric oxide synthase. *Proc Natl Acad Sci USA* **99,** 17167–17172 (2002).
62. Kirsch, T. *et al.* Abstract 516: Knockdown of the Hypertension Associated Gene NOSTRIN Alters Glomerular Barrier Function in Zebrafish (Danio rerio). *Hypertension* **60,** A516–A516 (2012).
63. XU, W., LIU, L. Z., LOIZIDOU, M., AHMED, M. & CHARLES, I. G. The role of nitric oxide in cancer. *Cell Research* **12,** 311–320 (2002).
64. Boyle, P. & levin, B. World Cancer report 2008. 1–260 (2008).
65. Bowtell, D. D. L. The genesis and evolution of high-grade serous ovarian cancer. *Nature Reviews Cancer* **10,** 803–808 (2010).
66. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502,** 333–339 (2013).
67. Nonaka, D., Chiriboga, L. & Soslow, R. A. Expression of Pax8 as a Useful Marker in Distinguishing Ovarian Carcinomas From Mammary Carcinomas. *The American Journal of Surgical Pathology* **32,** 1566–1571 (2008).
68. Hall, C. A. *et al.* Hippo Pathway Effector Yap Is an Ovarian Cancer Oncogene. *Cancer Res* **70,** 8517–8525 (2010).
69. Oji, Y. *et al.* Expression of the Wilms' Tumor Gene WT1 in Solid Tumors and Its Involvement in Tumor Cell Growth. *Cancer Science* **90,** 194–204 (1999).
70. Barbolina, M. V., Adley, B. P., Shea, L. D. & Stack, M. S. Wilms tumor gene protein 1 is associated with ovarian cancer metastasis and modulates cell invasion. *Cancer* **112,** 1632–1641 (2008).
71. Hylander, B. *et al.* Expression of Wilms tumor gene (WT1) in epithelial ovarian cancer. *Gynecologic Oncology* **101,** 12–17 (2006).
72. Schnitt, S. J. Classification and prognosis of invasive breast cancer: from morphology to molecular taxonomy. *Modern Pathology* **23,** S60–S64 (2010).
73. Becker, S. A historic and scientific review of breast cancer: The next global healthcare challenge. *International Journal of Gynecology & Obstetrics* **131,** S36–S39 (2015).
74. Malhotra, G. K., Zhao, X., Band, H. & Band, V. Histological, molecular and functional subtypes of breast cancers. *Cancer Biology & Therapy* **10,** 955–960 (2010).
75. Burdall, S. E., Hanby, A. M., Lansdown, M. R. & Speirs, V. Breast cancer cell lines: friend or foe? *Breast Cancer Res.* **5,** 89 (2003).
76. Attard, G. *et al.* Prostate cancer. *Lancet* **387,** 70–82 (2016).
77. David Cunningham, Z. Y. In vitro and in vivo model systems used in prostate cancer research. *Journal of biological methods* **2,** 17 (2015).
78. Stone, K. R., Mickey, D. D., Wunderli, H., Mickey, G. H. & Paulson, D. F. Isolation of a human prostate carcinoma cell line (DU 145). *Int. J. Cancer* **21,** 274–281 (1978).
79. Kaighn, M. E., Narayan, K. S., Ohnuki, Y., Lechner, J. F. & Jones, L. W. Establishment and characterization of a human prostatic carcinoma cell line (PC-3). *Invest Urol* **17,** 16–23 (1979).
80. Horoszewicz, J. *et al.* The LNCaP cell line--a new model for studies on human prostatic carcinoma. - PubMed - NCBI. *Prog Clin Biol Res* (1980).
81. Mertz, K. D. *et al.* Molecular Characterization of TMPRSS2-ERG Gene Fusion in the NCI-H660 Prostate Cancer Cell Line: A New Perspective for an Old Model. *Neoplasia* **9,** 200–IN3 (2007).
82. Sfanos, K. S. & De Marzo, A. M. Prostate cancer and inflammation: the evidence.

Histopathology **60,** 199–215 (2012).
83. Cavanagh, H. & Rogers, K. M. A. The role of BRCA1 and BRCA2 mutations in prostate, pancreatic and stomach cancers. *Hereditary Cancer in Clinical Practice 2015 13:1* **13,** 16 (2015).
84. Fidler, I. J. Tumor Heterogeneity and the Biology of Cancer Invasion and Metastasis. *Cancer Res* **38,** 2651–2660 (1978).
85. Allart-Vorelli, P., Porro, B., Baguet, F., Michel, A. & Cousson-Gélie, F. Haematological cancer and quality of life: a systematic literature review. *Blood Cancer Journal* **5,** e305 (2015).
86. Greaves, M. Leukaemia 'firsts' in cancer research and treatment. *Nature Reviews Cancer* **16,** 163–172 (2016).
87. Lenz, G. & Staudt, L. M. Aggressive Lymphomas. *N Engl J Med* **362,** 1417–1429 (2010).
88. Palumbo, A. & Anderson, K. Multiple Myeloma. *N Engl J Med* **364,** 1046–1060 (2011).
89. Mirfazaelian, H., Arbab, M. & Daneshbod, Y. Burkitt's lymphoma. *BMJ* **351,** h4545 (2015).
90. Leslie, K. K. *et al.* Endometrial Cancer. *Obstetrics and Gynecology Clinics of North America* **39,** 255–268 (2012).
91. Shang, Y. Molecular mechanisms of oestrogen and SERMs in endometrial carcinogenesis. *Nature Reviews Cancer* **6,** 360–368 (2006).
92. Hackenberg, R., Hawighorst, T., Hild, F. & Schulz, K.-D. Establishment of new epithelial carcinoma cell lines by blocking monolayer formation. *J Cancer Res Clin Oncol* **123,** 669–673 (1997).
93. Karnezis, A. N., Cho, K. R., Gilks, C. B., Pearce, C. L. & Huntsman, D. G. The disparate origins of ovarian cancers: pathogenesis and prevention strategies. *Nature Reviews Cancer* **17,** 65–74 (2016).
94. Zhou, X., Wang, Z., Zhao, Y., Podratz, K. & Jiang, S. Characterization of sixteen endometrial cancer cell lines. *Cancer Res* **67,** 3870–3870 (2007).
95. Zang, H., Li, N., Pan, Y. & Hao, J. Identification of upstream transcription factors (TFs) for expression signature genes in breast cancer. *Gynecological Endocrinology* **33,** 193–198 (2016).
96. Hackenberg, R. *et al.* Androgen responsiveness of the new human endometrial cancer cell line MFE-296. *Int. J. Cancer* **57,** 117–122 (1994).
97. Clark, M. A., Fisher, C., Judson, I. & Thomas, J. M. Soft-Tissue Sarcomas in Adults. *http://dx.doi.org/10.1056/NEJMra041866* **353,** 701–711 (2009).
98. Kansara, M., Teng, M. W., Smyth, M. J. & Thomas, D. M. Translational biology of osteosarcoma. *Nature Reviews Cancer* **14,** 722–735 (2014).
99. Mohseny, A. B. *et al.* Functional characterization of osteosarcoma cell lines provides representative models to study the human disease. *Laboratory Investigation* **91,** 1195–1205 (2011).
100. Visvader, J. E. Cells of origin in cancer. *Nature* **469,** 314–322 (2011).
101. Klein, A. M. *et al.* Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell* **161,** 1187–1201 (2015).
102. Stella, G. M., Senetta, R., Cassenti, A., Ronco, M. & Cassoni, P. Cancers of unknown primary origin: current perspectives and future therapeutic strategies. *Journal of Translational Medicine 2012 10:1* **10,** 12 (2012).

103. Hemminki, K., Riihimäki, M., Sundquist, K. & Hemminki, A. Site-specific survival rates for cancer of unknown primary according to location of metastases. *Int. J. Cancer* **133,** 182–189 (2013).
104. Pavlidis, N. & Pentheroudakis, G. Cancer of unknown primary site. *The Lancet* **379,** 1428–1435 (2012).
105. Kurahashi, I. *et al.* A Microarray-Based Gene Expression Analysis to Identify Diagnostic Biomarkers for Unknown Primary Cancer. *PLoS ONE* **8,** e63249 (2013).
106. Calder, K. B. & Smoller, B. R. New Insights Into Merkel Cell Carcinoma. *Advances in Anatomic Pathology* **17,** 155 (2010).
107. Feng, H., Shuda, M., Chang, Y. & Moore, P. S. Clonal integration of a polyomavirus in human Merkel cell carcinoma. *Science* **319,** 1096–1100 (2008).
108. Van Keymeulen, A. *et al.* Epidermal progenitors give rise to Merkel cells during embryonic development and adult homeostasis. *The Journal of Cell Biology* **187,** 91–100 (2009).
109. Harms, P. W. *et al.* Distinct Gene Expression Profiles of Viral- and Nonviral-Associated Merkel Cell Carcinoma Revealed by Transcriptome Analysis. *Journal of Investigative Dermatology* **133,** 936–945 (2013).

Appendix – Cancer CellNet code

```
#Cancer CellNet code - training and query
#Load all required files in S3

#ssh into the instance!!!
#configure your instance
aws configure

#create a folder in the ephemeral drive to accommodate the multiple files
sudo mkdir /media/ephemeral0/data
sudo chown ec2-user /media/ephemeral0/data
cd /media/ephemeral0/data

#copy the following files from S3 into the instance
#contains function definitions
aws s3 cp s3://cahanlab/pavithra.kumar/code/cellnetr_utils.R ./
aws s3 cp s3://cahanlab/patrick.cahan/projects/cancer.cellnet/stCancer_Sep_16_2016.R ./
#Training sample table
aws s3 cp s3://cahanlab/pavithra.kumar/code/CCLE_CellLines.csv ./
#Query sample table

sudo R
library(devtools)
install.github("pcahan1/CellNet") #install the Master CellNet package
q()

R
library(CellNet)
source("cellnetr_utils.R")
library(GEOquery)
#Bioclite analysis tools loaded
source("http://bioconductor.org/biocLite.R")
biocLite("affy")
biocLite("hgu133plus2.db")
biocLite("hgu133plus2cdf")
library("hgu133plus2cdf")
library("hgu133plus2.db")
library("affy")

sampTab <- read.csv("stCancer_Sep_16_2016.R") #load the training sample table
sids <- unique(as.vector(sampTab$exp_id))
row.names(sampTab) <- sampTab$sample_id

#Function definitons:
#1. Fetching files from S3, in parallel
s3_get_par <- function (bucket, path, fnames){
  tfname <- 'tmpfile.txt';
  write.table(fnames,tfname, col.names=FALSE, row.names=FALSE, quote=FALSE);
```

```
  cmd <- paste("cat ", tfname, " | parallel aws s3 cp s3://", bucket, "/", path, "/{} ./",
sep='');
  system(cmd);
}

#2. gzip the files
utils_unpack<-function (fname){
  cmd<-paste("gzip -d ", fname, sep='');
  system(cmd);
  fname<-strsplit(fname, ".gz")[[1]][1];
}

#3. Save the files in S3
s3_put<-function(dir, target, bucket="pcahanrnaseq"){
  fpath<-paste(bucket, "/",dir,sep='');
  cmd<-paste("aws s3 cp ",target," s3://", bucket, "/", dir, "/", sep='');
  cat(cmd,"\n");
  system(cmd);
}

#4. GRN construction
cn_make_grn <- function (sampTab, expDat, species = "Mm", tfs = NA, grnSampSize = 0,
    normDat = FALSE, corrs = NA, zscores = NA, cval = 0.5, cvalGK = 0.75,
    dLevel = "description1", dLevelGK = "description6", zThresh = 4,
    holmSpec = 1e-06) {
    targetGenes <- rownames(expDat)
    if (is.na(tfs)) {
        cat("Defining transcriptional regulators...\n")
        tfs <- find_tfs(species)
    }
    tfs <- intersect(targetGenes, tfs)
    if (grnSampSize == 0) {
        grnSampSize <- min(table(sampTab[, dLevel]))
    }
    stGRN <- sample_profiles_grn(sampTab, minNum = grnSampSize)
    cat("Number of samples per CT: ", mean(table(stGRN[, dLevel])),
        "\n")
    expGRN <- expDat[, rownames(stGRN)]
    if (normDat) {
        cat("Normalizing expression data...\n")
        expGRN <- Norm_quantNorm(expGRN)
    }
    if (is.na(corrs)) {
        cat("Calculating correlation...\n")
        corrs <- grn_corr_round(expGRN)
    }
    if (is.na(zscores)) {
        cat("Calculating context dependent zscores...\n")
        zscores <- grn_zscores(corrs, tfs)
    }
    grnall <- cn_getRawGRN(zscores, corrs, targetGenes, zThresh = zThresh)
    specGenes <- cn_specGenesAll(expGRN, stGRN, holm = holmSpec,
```

```
      cval = cval, cvalGK = cvalGK, dLevel = dLevel, dLevelGK = dLevelGK)
   ctGRNs <- cn_specGRNs(grnall, specGenes)
   list(overallGRN = grnall, specGenes = specGenes, ctGRNs = ctGRNs,
      grnSamples = rownames(stGRN))
}

#5. Heatmap - color annotations - definition
newHm <- function
### heatmap of the classification result
(cnRes,
 ### cellnet result
 isBig=FALSE
 ### is this a big heatmap
){
  classMat<-cnRes$classRes
  ## SORT BY TISSUE TYPE
  classMat = classMat[,order(colnames(classMat))]

  ## map color to tissue type
  columns = unique(colnames(classMat))
  counts = list()
  for(i in 1:length(columns)) {
    counts[[i]] = length(which(colnames(classMat) == columns[[i]]))
  }
  annotation_col = data.frame(TissueType = factor(rep(columns, counts)))
  ann_colors = list(c(columns = rainbow(length(columns))))
  ##
  colnames(classMat)  =  rownames(annotation_col)

  cools<-colorRampPalette(c("black", "limegreen", "yellow"))( 100 )
  bcol<-'white';
  if(isBig){
    bcol<-NA;
  }

  # generate heatmap
  pheatmap(classMat,
          col=cools,
          border_color=bcol,
          cluster_rows = FALSE,
          cluster_cols = FALSE,
          show_colnames = FALSE,
          annotation_col = annotation_col,
          annotation_colors = ann_colors)

}

#Begin Pre-processing!
#Fetch, decompress, and process by experiment id (sid)
y=1
mydate <- utils_myDate()
master_expQuery <- data.frame()
```

```
#Looping through samples from each experimental ID to get normalized expression data

for(sid in sids){
        stTest<-sampTab[which(sampTab$exp_id==sid), ]

   #downloading raw files from S3
        for(i in 1:nrow(stTest)){
                filename<- stTest$sample_id[i]
           tmpfname<-paste(filename,".CEL.gz", sep="")
           s3_get_par("pcahan_cn_cancer",
"data/human/array/hgu133plus2/raw_data/tumor_CEL_files", tmpfname)
           stTest$file_name[i] <- paste(stTest$sample_id[i],".CEL",sep="")
        }

        files <- list.files(pattern = "\\.CEL.gz$")
        for(ff in files){
                utils_unpack(ff)  #unpacking files
        }

        write.csv(stTest, file="expQ_expn.csv")

        stQuery<-expr_readSampTab("CCLE_expn.csv")

        expQuery<-Norm_cleanPropRaw(stQuery, "hgu133plus2")
        fname<-paste("expQuery_training_",sid,"_" mydate,".rda", sep='')
        save(expQuery, file=fname)
        s3_put("pavithra.kumar/CCN_runs/expQuery", fname, bucket="cahanlab") #save
the normalized exp matrix in S3

        master_expQuery <- cbind(master_expQuery, expQuery) #bind expQuery for all exp
ids to get master expQuery table

        #remove .CEL files to free space
        system(".rm *.CEL")

        #counter
        print(y)
        y <- y +1
}

save(master_expQuery, file = "master_expQuery_training.rda")
s3_put("pavithra.kumar/CCN_runs/expQuery",
"master_expQuery_training.rda",bucket="cahanlab") #save the master expQuery file


#GRN reconstruction

grnProp<- cn_make_grn(stAll, expAll, species = "Hs", tfs = hsTFs, dLevel =
"description1",
        dLevelGK = "description2", normDat=TRUE, cval = 0.5, cvalGK = 0.5)
```

```
fname<-paste("grnProp","_", mydate,".rda", sep='')
save(grnProp, file=fname)
s3_put("pavithra.kumar/CCN_runs", fname,  bucket="cahanlab")

#Split and assess

classifierPerformance <- cn_splitMakeAssess(stAll, expAll, grnProp, prop = 0.5, dLevel =
"description1",
        dLevelStudy = "exp_id", dLevelSID = "sample_id")
fname<-paste("classifierPerformance_new_cnProc","_",mydate,".rda", sep='')
save(classifierPerformance, file=fname)
s3_put("pavithra.kumar/CCN_runs", fname,  bucket="cahanlab")

cp <- classifierPerformance


#Classification heatmap
st <- sampTab[order(sampTab$description1),]
classMat <- cp$classRes
x <- intersect(rownames(sampTab),colnames(classMat))
classMat = classMat[order(rownames(classMat)),x]

cools<-colorRampPalette(c("black", "limegreen", "yellow"))( 100 )
bcol<-NA;

fname <- "Classifiers_heatmap.pdf"
pdf(file = fname)
pheatmap(classMat, col=cools, border_color=bcol, breaks=seq(from=0, to=1,
length.out=100), cluster_rows = FALSE,
        cluster_cols = FALSE, show_colnames = FALSE, fontsize = 10)
dev.off()
s3_put("pavithra.kumar/CCN_runs", fname, "_", mydate, bucket="cahanlab")

#Plot PR curves

fname<-paste("PRcurves_new_cnProc.pdf", sep='')
pdf(file=fname)
plot_class_PRs(cp$PRs)
dev.off()
s3_put("pavithra.kumar/CCN_runs", fname, bucket="cahanlab")


#Make cnProc
cnProc<-cn_make_processor(expAll, stAll, grnProp)
fname <- ("cnProc_",mydate, ".rda",sep='')
save(cnProc, file=fname)
s3_put("pavithra.kumar/CCN_runs", fname, bucket="cahanlab")

#Querying CellNet

sampleTab <- read.csv("CCLE_CellLines.csv")
rownames(sampleTab)<-as.vector(sampleTab$sample_id)
```

```
sids <- unique(as.vector(sampleTab$description1))
cName <- "cell_line"
y <- 1

for (sid in sids){
        stTest<-sampTab[which(sampTab$description1==sid), ]

 for(i in 1:nrow(stTest)){
  filename<- stTest$sample_id[i]
          tmpfname<-paste(filename,".CEL.gz", sep="")
          s3_get_par("cahanlab", "pavithra.kumar/osteosarcoma_RAW", tmpfname)
          files <- list.files(pattern = "\\.CEL.gz$")

   for(ff in files){
            utils_unpack(ff);
           }
          stTest$file_name[i] <- paste(stTest$sample_id[i],".CEL",sep="")
         }

         write.csv(stTest, file="CCLE_expn.csv")

         stQuery<-expr_readSampTab("CCLE_expn.csv")

         expQuery<-Norm_cleanPropRaw(stQuery, "hgu133plus2")

         fname<-paste("expQuery_osteosarcoma_", mydate,".rda", sep='')
         save(expQuery, file=fname)
         s3_put("pavithra.kumar/CCN_runs/expQuery", fname, bucket="cahanlab")

         cnObjName<-(hgu133plus2 = cnProc)

 tmpAns<-cn_apply(expQuery, stQuery, cnProc, dLevelQuery=cName);

         fname<-paste("tmpAns_", sid, mydate,".rda", sep='')
         save(tmpAns, file=fname)
         s3_put("pavithra.kumar/CCN_runs/tmpAns", fname, bucket="cahanlab")

 fname<-paste("HeatMap_", sid,"_", mydate,".pdf", sep='')
 pdf(file=fname, width=40, height=40)
 newHm(tmpAns, isBig=TRUE)
 dev.off()
 s3_put("pavithra.kumar/CCN_runs/Output", fname, bucket="cahanlab")

         system("rm *.CEL")

         print(y)
         y<-y+1
}
```

Vita

Pavithra Kumar received her B. Tech in Bioengineering from SASTRA University, India in 2012 and has gained research experience in Harvard University, Indian Institute of Science and TIFR, Mumbai, before coming back to the US to pursue her Masters in Biomedical Engineering in 2015. Her broad research interests include cancer diagnostics and personalized medicine. She joined the lab in August 2015 and her initial work focused on setting up a microfluidic based single cell capturing platform called Dropseq, that could be used to obtain gene expression profiles at a single cell resolution. For her thesis, she modified CellNet, a computational platform, to assess the fidelity of cancer cell lines as in vitro tumor models by comparing their gene regulatory networks. She was born and raised in Chennai, India and in addition to Science, she enjoys dancing, trekking, eating chocolate and reading historical novels.